

Requirements for Data Validation to Support XML-Enabled Wide Area Search in Bioinformatics: A Position Paper

Judith D. Cohn

DOE Joint Genome Institute/Bioscience Division, Los Alamos National Lab

Keywords: XML, wide area search, bioinformatics, genomics, data validation

Introduction

Regardless of whether or not the biological community chooses to fully support a few tightly integrated, centralized databases or a larger number of distributed web sites with a common interface, users face a formidable barrier to the efficient use of the ever-growing stores of data. This barrier is the inconsistent quality of the available data and the problem of filtering out the “good” from the “bad” or merely unknown quality data.

In many ways the move towards a system of distributed web sites rather than monolithic central databases makes this problem easier. Instead of relying on the owners of a few large databases to offer the appropriate quality information directly, data validation or quality filtering could be offered as a separate distributed service. Thus, while Organizations A, B and C could offer data access services; Organizations C, D and E could offer data validation services, perhaps of differing stringencies appropriate for a variety of applications. Note that it is not required that the data access service offer data validation or vice versa.

Learning from eCommerce

In the business world, two different paradigms for eCommerce transactions are emerging: B2C and B2B. B2C (Business to Consumer) transactions require a web site (the business or B) at one end and an interactive consumer (C) pointing and clicking at the other end. An example would be a consumer buying books from Amazon.com. B2B or Business to Business transactions require only a computer at each end without the need for human intervention. Thus, General Motors or Ford might interact with individual vendors or a public trading exchange to obtain the parts necessary to build an automobile. Wide area searches in bioinformatics most closely resemble the B2B paradigm.

In the B2B world, there are two categories of data validation: the who and the what. The “who” part is the need to validate whether or not a contract exists between the two businesses (i.e. is company Y an authorized tire vendor for General Motors) and then to confirm that the participating computers actually are authorized to represent the appropriate companies. This is to a large extent a security issue. The “what” part is the need to validate whether or not the offered product or service meets the specifications of the current request. This is a domain issue - i.e. both purchasers and vendors have to agree on what comprises a specification and what comprises a match.

Distributed Bioinformatics Data Validation

In the bioinformatics world, we can also divide data validation into the who and the what. The who problem is much simpler to solve in the short term than the what problem – at least *vis a vis* the publicly available databases, which currently do not charge a fee for access. The what problem, which involves agreeing to a relatively complex level of XML tags, DTDs, etc. to support the wide variety of quality measures involved in biological research, is more thorny and will take longer to solve. It, therefore, might prove useful to first concentrate on the who part, which might be implemented while we continue to work on the what problem.

One very simple way to tackle the who problem would be for various organizations to offer a list of data access sites, which have been “validated” for a specific type of data and a specified level of stringency. It would be up to users to determine which validation sites could be trusted. Under this type of scenario, instead of a sequencing center submitting data to Genbank by physically sending the data, Genbank would validate the data (by whatever means it chose), giving the web site a rating, which end users could take into account when doing data searches. These ratings could be obtained by sending a query to the validation service. The ratings would be per database or data access service. This might serve as an incentive for organizations to separate out the wheat from the chaff in their data – perhaps offering different versions or pieces of their databases as a separate data access service, each obtaining a separate rating.

Another way that a validation service could help filter data for end users would be to serve as the search engine itself. In a similar fashion to the way that web users might choose altavista over excite to find information on a particular subject, biologists might turn to data validation service X over validation service Y for high quality gene expression data. In this way, biologists would make queries to their validation service of choice, using the generic search query standards agreed upon by the biological community; and the service itself would selectively search appropriate data bases.

