Markup of microbiological data for accelerated publication in print and electronic form.

George M. Garrity
Bergey's Manual Trust/Department of Microbiology

While in its infancy, the field of bioinformatics is rapidly emerging and will be an essential area of research for the foreseeable future. Spurred on by rapid advances in laboratory instrumentation, it is now possible to generate huge volumes of data at a constantly lowering price/data point. However, possession of large volumes of raw data is not enough. Such data must be subjected to rigorous and careful analysis; a cognitive activity that remains a rate limiting step. This requires a sound knowledge of the associated scientific literature and complementary databases. To date, bioinformatics research has focused predominantly on sequence data. However, the scope of the field must broaden to meet the growing needs of the community at large. Related biological information (physiological and structural) must be overlaid onto the sequence data to provide meaning. Accomplishing this next step will require a major change in the way in which we produce and report biological information.

Traditionally, the bulk of biological information has been delivered in the form of free text, using unconstrained, and at times, imprecise English. While this is a convenient way to communicate, it is exceedingly difficult to index and retrieve such content by machine. The problem is further exacerbated by the sheer volume of new information published annually. It would be highly advantageous if tools or services were available to permit better access to the biological literature. Ideally, natural language processing would be the best solution, but this approach has not yet been successful. There are, however; alternative approaches.

Standard Generalized Markup Language (SGML) was developed in the 1980's by IBM as a tool for expediting the production of highly structured technical documentation. The goal was to make such content more manageable, reusable and portable. SGML provides a means of identifying and separating textual elements, which are content specific, from typesetting commands which are hardware specific and appear inline. Over that past decade, SGML has been widely adopted as a means of production and management of technical documentation in the automotive, aerospace, telecommunications and computer industries. More recently, some scientific publishers have begun exploring the potential of this technology, however, the initial investment in both special software and technical expertise is high.

Although SGML is the source language of HTML (and XML), both SGML and XML offer significant advantages and flexibility in content management and presentation. However, a prerequisite to the use of these languages is a preliminary declaration of all of the tags that will be used to identifying textual and graphic elements. This information is contained within the Document Type Definition that accompanies the content and defines all of the rules of tag usage. While DTD creation requires a large, up-front investment, the payoff is the ability to search and modify extremely large bodies of free text based on the identifying tags. This significantly reduces the number of spurious and irrelevant hits one encounters using alternative tools.

Bergey's Manual Trust is a non-profit educational trust, housed in the Microbiology Department at MSU. We are the leading provider of reference books in the field of bacterial systematics and are engaged in the production of a new edition of *Bergey's Manual of Systematic Bacteriology*. The *Manual* is an encyclopedic treatment of all the known bacteria and will include approximately 1350 chapters, authored by approximately 650 international experts. After several failed attempts at using conventional databases for content management, we have adopted SGML as core technology. Over the past six months, we have developed an extensible DTD that will accommodate all of the content used in our publication. Explicit in our design is the ability to extend this DTD to include a rich set of tags for genetic, physiological and structural characteristics of bacteria with variable levels of granularity. This discussion will focus on DTD design considerations and high-end commercial tools (ArborText Adept Editor and Chrystal Document Management System) used on this project.