

BSML: Origins, Overview and Status

BSML™ (Bioinformatic Sequence Markup Language) is an open XML data representation and interchange format that unifies diverse technologies and enables more efficient communication within the life sciences community. BSML, the first XML application in the life sciences, was developed under a 1997 grant from the National Human Genome Research Institute (NHGRI) to provide standard methods for communicating genomic research information. Under the direction of Joseph Spitzner, Ph.D., now Rescentris' Chief Technical Officer, BSML was created as an evolving public domain standard.

BSML was developed to allow life sciences researchers, suppliers, and content and service providers to interact and exchange information through a universal data language that captures the richness of bioinformatics and genome research data. Through the use of XML encoding, BSML documents retain the biological meanings and relationships of the underlying information. BSML describes biological information, including sequence data, feature tables, literature references, and associated tabular data (e.g., gene expression values).

The latest release, BSML 3.1, provides content models for describing various types of research — queries, searches, analyses, and experiments. Research descriptions may be associated with sequence and other annotations to provide a complete description of the source of evidence for a claim. BSML also encodes display information, such as display widgets, custom viewers, and interaction controls for bioinformatic content. BSML uses logical page views and collections of pages for presentation — a notebook format for biologists that supports linking, metadata, and annotations of all the individual objects on a page. Note that BSML documents do not need to contain display information and may be completely valid and well-formed XML using only the data encoding and/or research portions of BSML.

Contributors.-- Following the initial RFC, BSML evolved from 1997 to 1999 with input from various people (e.g., Lincoln Stein at Cold Spring Harbor, Peter Murray-Rust, codeveloper of Chemical Markup Language). BSML 2.2 was released in 2000. After that, many improvements were considered, with input from a variety of governmental, academic, and commercial sources, including the staff at LabBook, Inc., and members of the I3C. The result of this process was **BSML 3.1**, completed in June 2002.

BSML in the public domain -- From the start, BSML has been released to the common domain. No licensure agreement or license fee is needed to make use of BSML for any purpose, commercial or otherwise, providing that copyrights are acknowledged. In addition to the BSML 3.1 DTD, LabBook and Rescentris have also made available detailed reference and tutorial manuals (also authored by Joseph Spitzner). The copyright to BSML is held by LabBook, Inc. Use of the BSML trademark, DTD, and technical documents have been licensed to Rescentris, Ltd. These documents may be freely distributed as long as they are not modified and copyrights are acknowledged. More information about BSML can be obtained at <http://www.rescentris.com> and at <http://www.bsml.org>.

Becoming a public standard.-- The initial motivation for BSML remains today -- to create a useful public domain standard for the representation, integration and communication of bioinformatic content. To become a *useful data representation and exchange standard*, BSML must meet three objectives:

- Achieve **interoperability** through platform independent usability and interconvertability with other data representation formats
- Gain **endorsement** from the diverse community of users
- Go through a **formal standardization process** administered by appropriate standardization bodies

Interoperability.-- BSML is an XML application. Like all XML, it is inherently platform independent. Implementations of BSML have used several programming languages (Java, Delphi/Pascal, Perl), which have been implemented on a variety of operating systems and hardware platforms (Windows, Mac, Linux, Unix). BSML has been converted from other representations using Perl scripts that are publicly available through BioPerl (courtesy of Bristol-Myers Squibb). BSML has also been converted from the NCBI ASN.1 representation of GenBank and from a variety of flatfile database representations (GenBank, EMBL, Ensembl, DDBJ, OHGD, and Swiss-Prot). Most of these conversions are freely available using Rescentris' Genomic Workspace™ Viewer and other software from Rescentris (www.rescentris.com). Also see www.bsml.org.

Endorsement.-- BSML has been supported and used by a growing number of organizations; a partial list includes Accelrys, ApoCom Genomics, Biotechnology Industry Organization (BIO), Bristol-Myers Squibb, EMBL's European Bioinformatics Institute (EBI), Fujitsu, IBM, Itochu, I3C, John Wiley & Sons, National Foundation for Cancer Research, NetGenics., OhioSupercomputer Center, and The Institute for Genome Research (TIGR).

Formal standardization process.-- BSML 3.1 is being released as a final draft for comment through the American Society for Testing and Materials (ASTM Healthcare Informatics, Subcommittee for Bioinformatics), following a meeting in Seattle, on May 12, 2002. Through ASTM's relationship with ANSI, it is expected that endorsement of BSML by ASTM will be rapidly followed by endorsement from ANSI. The BSML 3.1 specification has also been provided for members of the I3C (Interoperable Informatics Infrastructure Consortium), some of whom have been reviewing BSML for a considerable period of time. It is expected that I3C members will have significant input to the review process of ASTM. BSML 3.1 will also be provided to CENSA (Collaborative Electronic Notebook Systems Association) as a life sciences data model and as a guideline for development of new standards for collaborative e-R&D.

BSML Features

BSML is composed of three sections:

- **Definitions** – encoding of genomes and sequences, data tables, sets, and networks
- **Research** - encoding of queries, searches, analyses and experiments
- **Display** – encoding of display widgets that represent graphical representations of biological objects

BSML 3.1 contains a number of improvements and new content over the previously released version 2.2:

Genomics encoding.-- BSML 3.1 encodes genomic information from the level of the complete genome down to the base pair. A complete hierarchy is available for genomes (**Cell-line - Genome - Chromosome - Cytoband**) and the sequences they contain (**Sequence - Feature-table - Feature - Qualifier - Modification**). Through the use of relational elements (**Alignment, Segment-set**, etc.), it is possible to encode inter-sequence relationships (alignments, homologies, genes and their products).

Isoforms.-- The encoding of isoform related content is new to BSML 3.1. This section includes representation of mutations, SNPs, phenotypes, genotypes and pedigrees. The **Case** element allows individual organisms to be described in terms of **Phenotype** and **Isoform** changes in biological sequences.

Networks and pathways.-- Also new to BSML 3.1 is the representation of networks through graph elements (**Node** and **Arc**). This encoding allows the description of metabolic pathways along with other uses of network representation (e.g., gene expression).

Research encoding.-- A new high level section of BSML 3.1 encodes descriptions of **Research**. Elements are available for **Queries, Searches, Analyses** and **Experiments**, including full descriptions of research **Protocols**. Analytical results may be encoded in elements available from earlier versions of BSML (e.g., **Table-import**) or using some of the new elements provided with BSML 3.1 (e.g., **Sequence-search-table**). The combination of sequence description, research description, and data table encoding in BSML allows a complete representation for many analyses.

Metadata, ontologies, and controlled vocabularies.-- One of the needs in knowledge representation is the ability to make statements about data. Several standard approaches to such metadata have emerged in recent years, and BSML 3.1 provides a complete representation of the Dublin Core metadata standard. This standard includes statements about such attributes as authorship, subject, version, etc., and permits straightforward description of bibliographic content, patent citations, and database builds.

The metadata elements associated with the BSML 3.1 **Resource** element provide other functionality. For example, the source of any annotation of sequence related features may be clearly identified using the **Resource** element.

One of the functions of metadata is to provide explicit frames of reference for terminology. BSML 3.1 provides a set of **Authority** elements that indicate the source of a term and permit access to additional information about it. This powerful feature allows users to specify the meaning of their terms as required by controlled vocabularies and ontologies.

Representing relationships through XML linking and cross-references.-- The representation of biological objects is of little use unless the relationships among objects can be expressed. BSML 3.1 provides a full range of XML linking elements, including a **Cross-reference** element that allows encoding of the type of relationship that links two objects (e.g., x is a gene-product of y). This type of linking allows the representation of all types of relationship (homology, expression, etc.) and allows simple navigation among BSML documents and to other documents. It is this feature that provides some of the knowledge integration power of BSML. Through linking, BSML can integrated not only the content types that it encodes directly (e.g., a restriction site table), but also other data types that may be represented in external documents (e.g., an expression table or NCBI BLAST result set).