

BioXSD: the *canonical XML-Schema data model* for everyday bioinformatics Web services

Matúš Kalaš^{1,2}, Pål Puntervoll¹, Alexandre Joseph³, Edita Bartaševičiūtė⁴, Armin Töpfer^{1,5}, Jon Ison⁶,
Christophe Blanchet³, Kristoffer Rapacki⁴ and Inge Jonassen^{1,2}

¹Computational Biology Unit, Bergen Center for Computational Science, Uni Research and ²Department of Informatics, University of Bergen, Bergen, Norway. ³Université Lyon 1; CNRS, UMR 5086; IBCP, Institut de Biologie et Chimie des Protéines, 69367 Lyon Cedex 07, France. ⁴Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark. ⁵Institute for Bioinformatics, Center for Biotechnology, Bielefeld University, Bielefeld, Germany. ⁶European Bioinformatics Institute, EMBL, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

Abstract

The wide community of life-scientific groups offers a huge amount of public bioinformatics resources and tools. These resources cover diverse areas of bioinformatics and follow diverse sets of implementation designs. More and more of the resources provide a standardised Web-service interface. But we are witnessing a burden to smooth interoperability among the services: a lack of standard input and output data formats.

Although the flat-file textual or tabular formats have played an important role in bioinformatics, the interest in machine-friendly data formats and interoperable Web services makes the use of XML highly advantageous. *XML Schema* (or XML-Schema Document, XSD) formally defines the structure of the data consumed or produced by the Web services. This formal definition of the data format is machine-readable, and there are well-proven industrial technologies for processing the XML Schema and the XML data.

A fine-grained XSD definition allows automatic validation of the data and lets input data be 'parsed on arrival' without any need for proprietary parsers. It enables straight-forward translations into different data formats. The fine-grained components of the data types can be semantically annotated by terms from a controlled vocabulary/ontology for describing bioinformatic data, such as the EMBRACE Ontology for Data and Methods (EDAM).

BioXSD, initiated by the EMBRACE project partners, offers a set of such fine-grained formal definitions of the basic input and output data formats. It attempts to serve as the common, standard data model for the most widely used biological data exchanged with Web services. *BioXSD* covers biological sequences, alignments, sequence annotations with both positional and non-positional features, and references to databases and ontologies. Defined *BioXSD* types are annotated by the EDAM ontology, which offers formalised, controlled meanings of the data types.

BioXSD has been developed, and is in the process of further development and refinement, by analysing the requirements of existing Web services, tools, ontologies and the various existing data formats, and in a wide and ongoing collaboration within the bioinformatic community. *BioXSD* allows users to mix-and-match diverse services freely and without a need for 'shims' to translate between the plenty of legacy data formats. The programming or design of analytical workflows becomes easier, faster, and cheaper, decreasing the needs for specialised personnel with advanced programming skills.

The service providers can use *BioXSD* directly when applicable, or develop custom types which can include the canonical ones, or further extend or restrict them. With services that use proprietary formats, *BioXSD* can be used as the canonical intermediate exchange format. For some specialised fields of bioinformatics, dedicated standard XML Schemas, or at least XML languages do exist. These include for example SBML, PDBML, or MAGE-ML. The specialised XSDs are optimally to be used whenever applicable, hand-in-hand with *BioXSD* which focuses on the most common, basic data-types, not yet globally standardised in XML. Transformers between the *BioXSD* and the main community textual or tabular formats are included in the *BioXSD* development.