BioXSD — a data model for sequences, alignments, features, measured and inferred values

Matúš Kalaš[1], Sveinung Gundersen[2], László Kaján[3], Jon Ison[4], Steve Pettifer[5], Christophe Blanchet[6], Rodrigo Lopez[7], Kristoffer Rapacki[4] and Inge Jonassen[1]

[1] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway.
Email: matus.kalas@uib.no
[2] Department of Informatics, University of Oslo, Oslo, Norway.
[3] unaffiliated (previously Bioinformatics and Computational Biology Department, Technische Universität München, Garching, Germany).
[4] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark.
[5] School of Computer Science, University of Manchester, Manchester, UK.
[6] French Institute of Bioinformatics, Gif-sur-Yvette, France.
[7] European Bioinformatics Institute, EMBL, Hinxton, UK.

Project Website: http://bioxsd.org
Source Code: https://github.com/bioxsd/bioxsd

Note: We have been exploring ways to adopt, adapt, or develop a license – or a combination of a license and some additional technical and ethical rules - that would be suitable for community-developed, "open" standards for interoperability. While openness for contributions, improvements, and certain kinds of customizations is one of the main goals, another main goal is keeping a standard "standardized" enough to serve the desired interoperability. We would like to work together with O|B|F on establishing such foundations suitable for interoperability standards developed in a participatory and transparent community spirit, and draw attention to **open development and licensing of standards for interoperability** during the BOSC Codefest and the BOSC 2015 itself.

BioXSD has been developed as a universal data model and exchange format for basic bioinformatics types of data: sequences, alignments, features and related values, inferred or measured. The BioXSD data model is rich enough to enable loss-less capture of diverse data that would otherwise require use of multiple different formats, and often even introduction of new formats for untypical features, classifications, or measured values. In BioXSD, an innovatively broad range of experimental data, annotations, and alignments can be recorded in an integrated chunk of data, together with provenance metadata, documentation, and semantic annotation with concepts from ontologies of user's choice.

BioXSD has so far been released in form of a machine-understandable XML Schema (XSD). Ongoing developments concentrate on providing BioXSD in form of JSON Schema and XML Schema 1.1, which may in the future be supplemented by RelaxNG, or even OWL and other data-modelling languages or frameworks. This will enable using BioXSD as a common data model supporting serialization of bioinformatics data into XML, JSON, RDF, or binary (EXI and BSON) as desired, while maintaining consistent and smooth validation, conversions, and parsing into objects for programming. The semantics of BioXSD is defined via SAWSDL references to EDAM (http://edamontology.org) and to a few main Semantic-Web vocabularies.