

BioXSD: the XML Schema for basic bioinformatics data

Matúš Kalaš^{1,2}, Pål Puntervoll¹, Edita Bartaševičiūtė³, Christophe Blanchet⁴, Armin Töpfer^{1,5}, Jan Christian Bryne⁶, Jon Ison⁷, Kristoffer Rapacki³, and Inge Jonassen^{1,2}

Contact: support@bioxsd.org

¹Computational Biology Unit, Uni Computing, Bergen, Norway. ²Department of Informatics, University of Bergen, Norway. ³Center for Biological Sequence Analysis, Technical University of Denmark, Kongens Lyngby, Denmark. ⁴Institut de Biologie et Chimie des Protéines, CNRS, Université Lyon 1, France. ⁵Institute for Bioinformatics, Bielefeld University, Germany. ⁶Institute for Cancer Research, Oslo University Hospital, Norway. ⁷European Bioinformatics Institute, EMBL, Hinxton, UK.

BioXSD.org

<http://bioxsd.org/BioXSD-1.1beta1.xsd>

Available under the **Creative Commons BY-ND 3.0** licence with additionally allowed inclusion, extensions and restrictions in user's XML namespace. Contributions to new canonical versions, in the *bioxsd.org* XML namespace, are welcome under supervision of the BioXSD consortium (in order to keep BioXSD a common, canonical data model).

A common XML Schema (XSD), defining a canonical XML format, is important for smooth compatibility of heterogeneous tools and data resources, and in particular for communication with Web services. The lack of a common XSD-based format for the basic bioinformatics types of data has motivated the development of BioXSD [1].

BioXSD is an XML Schema defining formats of the main bioinformatics types of data that are not modelled by any specialised standard XML Schemas such as for example SBML, PDBML, MAGE-ML, MIF, GCDML, or phyloXML [2-7]. BioXSD thus focuses on biomolecular sequences, alignments, and annotation by any kind of features or properties. These main types are accompanied by definitions of data-resource and ontology reference formats, provenance metadata, scores, and other. BioXSD has been coordinated with the European EMBRACE project focused on practical interoperability among bioinformatics tools [8].

The aim of BioXSD is to become used as a canonical, "standard" data format for sequence data and generic feature annotations. It does not mean that BioXSD should be "the only format", but an exchange format that can be common to several tools (as one of multiple formats the tools are supporting). Tools can produce and consume BioXSD directly, or BioXSD can be used as an intermediate canonical format rich enough to enable conversions among diverse formats. BioXSD types can be directly included into other XML Schemas, or they can be further extended or restricted in a similar way to object-oriented programming classes. The XML Schema can serve as a specification for generating more efficient binary representations such as EXI [9]. BioXSD formats

have detailed structure and are rich enough to support varying requirements for machine-understandable data and metadata representation, but at the same time trying not to be too complicated. Semantics of the syntactic BioXSD types is defined via SAWSDL annotation with concepts from the EDAM ontology [10].

The highlights of the BioXSD format itself are: structured metadata of simple sequence records (as opposed to FASTA *defines*), provenance metadata in all types of processed data and references, structured references to data resources and ontology concepts including a *meaning* of the relation, complex relations between sequence features, feature model data, multiple complex scores with meanings, formalised annotation of related positions outside of the annotated sequence, and more. The new BioXSD version 1.1 has optimised the syntax for feature annotation, scores, semantic and data references. It allows annotation of dense sequence features applicable to whole-genome annotations exchanged for example in the standardised binary EXI format. The first beta version of BioXSD 1.1 has been released in May 2011. Extensive support for whole-genome alignments, individual genomics, and sequence profiles will be added in the next beta versions.

In future, BioXSD must be maintained and regularly refined in order to fit the diverse and changing needs of the bioinformatics community. Involvement of the community is essential both for the uptake and the further development which must be coordinated by the emerging BioXSD consortium. To enable larger-scale adoption, supportive programmatic and interactive tools have to be developed, including format converters and integration with the O|B|F Bio* frameworks.

[1] Kalaš, M. *et al.* (2010) BioXSD: the common data-exchange format for everyday bioinformatics web services. *Bioinformatics*, **26**, i540-i546.

[2] Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524-531.

[3] Westbrook, J. *et al.* (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988-992.

[4] Spellman, P.T. *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, **3**, research0046.1-0046.9.

[5] Hermjakob, H. *et al.* (2004) The HUPO PSI's Molecular Interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177-183.

[6] Kottmann, R. *et al.* (2008) A standard MGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS*, **12**, 115-121.

[7] Han, M.V. and Zmasek, C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.

[8] Pettifer, S. *et al.* (2010) The EMBRACE Web service collection. *Nucleic Acids Res.*, **38**, W683-W688.

[9] Efficient XML Interchange (EXI) Format 1.0. <http://www.w3.org/TR/exi/>

[10] <http://edamontology.sourceforge.net>