BioXSD

A data model for sequences, alignments, features, and measurements

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

[1]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway; [2]Institute for Cancer Research, Oslo University Hospital and Department of Informatics, University of Oslo, Oslo, Norway; [3]unaffiliated, previously Bioinformatics and Computational Biology Department, Technische Universität München, Garching, Germany; [4]European Bioinformatics Institute, EMBL, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; [5]School of Computer Science, University of Manchester, Manchester, UK; [6]L'Institut Français de Bioinformatique, Gif-sur-Yvette, and Institut de Biologie et Chimie des Protéines, CNRS and Université Claude Bernard Lyon 1, Lyon, France; [7]Center for Biological Sequence Analysis, Technical University of Denmark, Kongens Lyngby, Denmark.

support@bioxsd.org

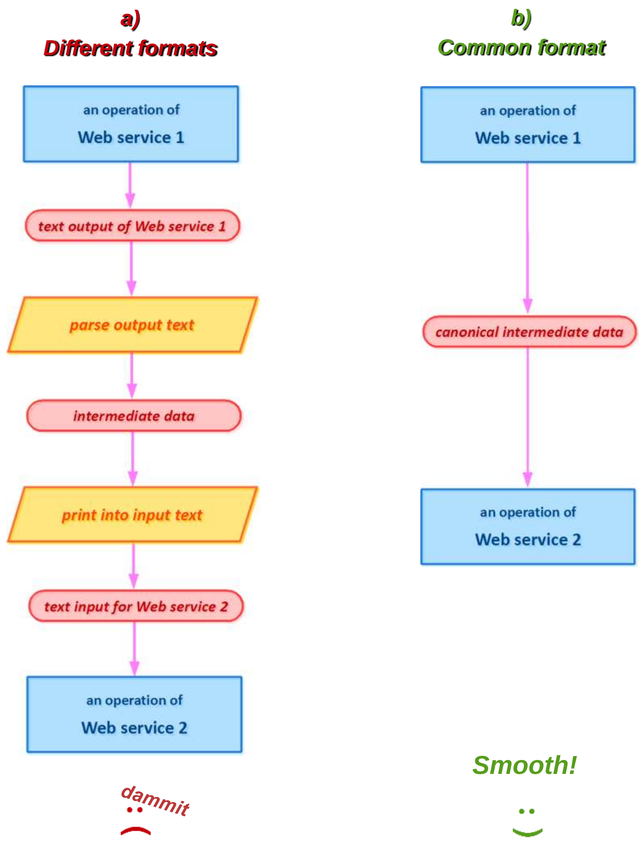
MOTIVATION

Without a common format, communication between diverse tools demands conversions, "shims", or do-it-yourself parsing. And worst of all, **maintaining** them in the future.

The 2 scenarios show demands for connecting 2 tools (for example web services) that use:

a) Different formats

b) Common format



TECHNOLOGY CHOICES

Different paradigms of data formatting represent data differently

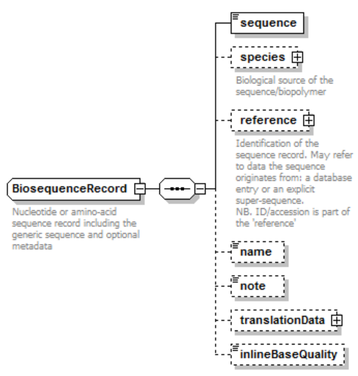
Traditional formats:	Middle way:	Semantic Web:
1D Key-value pairs	XML format	RDF format
2D Tab-separated values	JSON format	
Binary format	**Tree**	**Graph**

A machine-understandable definition of a specific format (a data model) is highly beneficial for validation and maintainability

GTrack format	XML Schema (XSD) Relax NG XML Schema 1.1 JSON Schema	OWL

EXAMPLE RESULT: BioXSD sequence record

Type diagram from the schema:

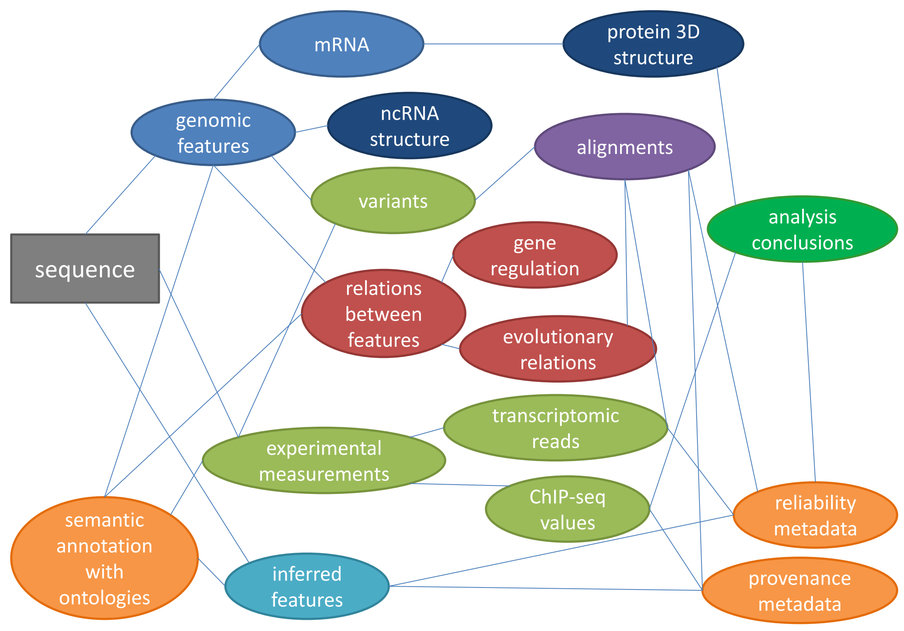


Example data instance:

```
<mySequence xsi:type="bx:GeneralAminoacidSequenceRecord">
    <bx:sequence>MDPLGDTLRRLREAFHAGRTRPAEFRAAQLQGLGRFLQENKQLLHDALAQDLHKSAFESEVSEVAISQGEVTL
AELGGKNPCYVDDNCDPQTVANRVAWFRYFNAGQTCVAPDYVLCSPEMQERLLPALQSTITRFYGDDPQSSPNLGRIINQKQFQRLRALLG
GKFSFDTFSHHRACLLRSPGMEKLNALRYPPQSPRRLRMLLVAMEAQGCSCTLL</bx:sequence>
    <bx:species
        dbName="NCBI Taxonomy"
        accession="9606"
        entryUri="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606"
        speciesName="Human"
    />
    <bx:reference
        dbName="UniProt"
        accession="P43353"
        entryUri="http://www.uniprot.org/uniprot/P43353"
        sequenceVersion="1"
        variantAccession="P43353-1"
    />
    <bx:name>Aldehyde dehydrogenase family 3 member B1 (ALDH3B1)</bx:name>
</mySequence>
```

EXAMPLE RESULT: BioXSD feature record

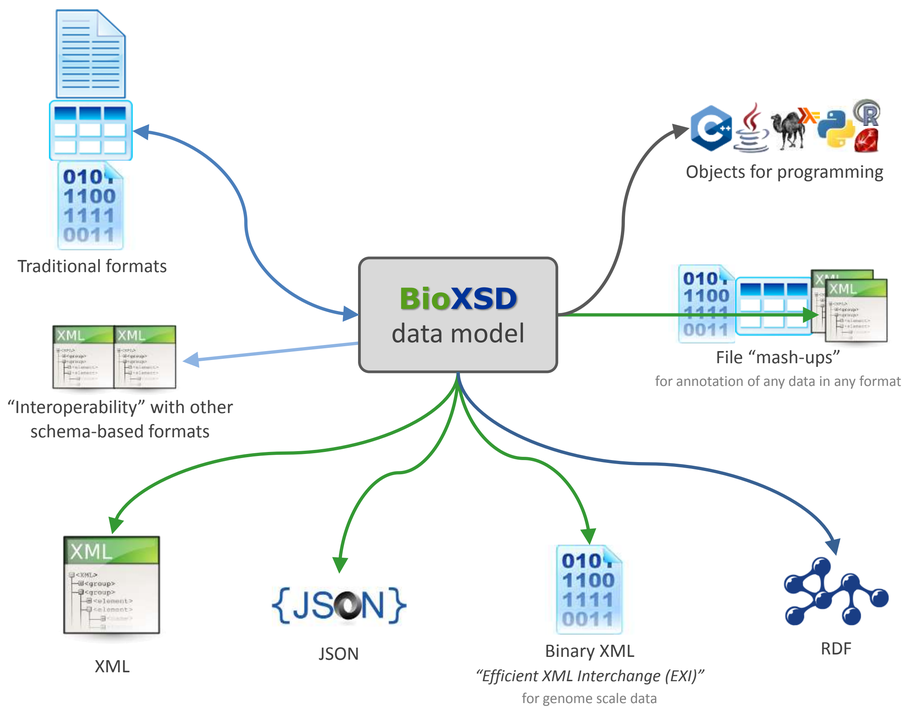
BioXSD can represent diverse interconnected data and annotations in an integrated data record



ONGOING DEVELOPMENT: single data model with multiple choices of exchange formats and conversions

BioXSD is rich enough to enable:

- loss-less capture of diverse data that would otherwise require use of multiple different formats
- loss-less conversions between diverse formats



BioXSD.org