BioXSD
BioJSON BioYAML

Towards unified formats for sequences, alignments, features, and annotations

Matúš Kalaš[1], Sveinung Gundersen[2], Inge Jonassen[1], and the BioXSD and GTrack contributors

[1]*Computational Biology Unit, Department of Informatics, University of Bergen, Norway;* [2]*Department of Informatics, University of Oslo, Norway;* ✉*developers@bioxsd.org*.

/bioxsd/bioxsd *@BioXSD* *http://groups.google.com/group/bioxsd* *http://bioxsd.org* ✉ *support@bioxsd.org*

Latest stable release: *http://bioxsd.org/BioXSD-1.1.xsd*

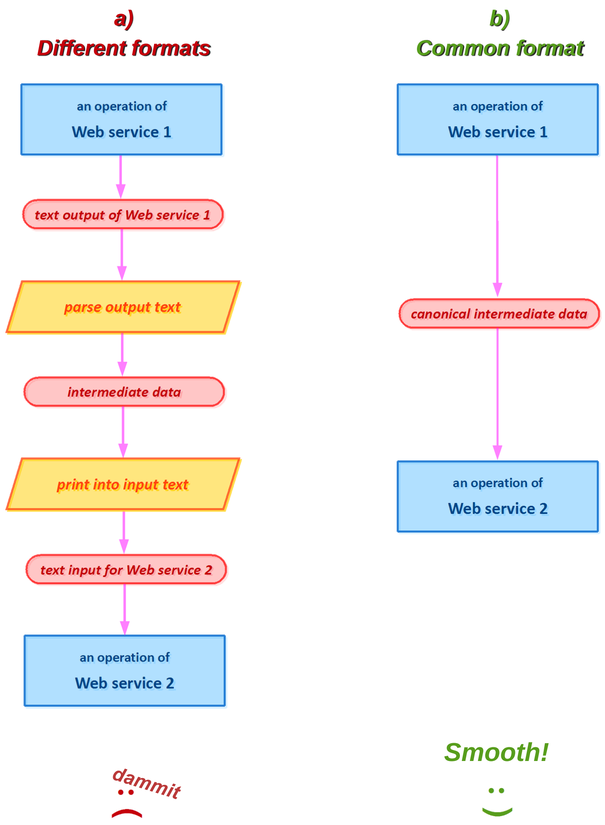
MOTIVATION

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

The 2 scenarios show demands for connecting 2 tools (e.g. Web services) that use:

a) Different formats

b) A common format



TECHNOLOGY CHOICES

Different paradigms of data formatting represent data differently.

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

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

GTrack format	XML Schema (XSD) 1.0 XML Schema 1.1 Relax NG JSON Schema ...	OWL ...
TSV with column definitions http://gtrack.no		

SIMPLE EXAMPLE: BioXSD sequence record

Example data instance, BioXSD in XML:

```
<mySequenceRecord
        xsi:type="bx:GeneralAminoacidSequenceRecord"
        xmlns:bx="http://bioxsd.org/BioXSD-1.1"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://bioxsd.org/BioXSD-1.1 http://bioxsd.org/BioXSD-1.1.xsd"
>
    <bx:sequence>MDPLGDTLRRLREAFHAGRTRPAEFRAAQLQGLGRFLQENKQLLHDAL</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:subsequencePosition>
            <bx:segment min="1" max="48"/>
        </bx:subsequencePosition>
    </bx:reference>
    <bx:name>Aldehyde dehydrogenase family 3 member B1 (ALDH3B1), N-terminus</bx:name>
</mySequenceRecord>
```

In BioJSON:

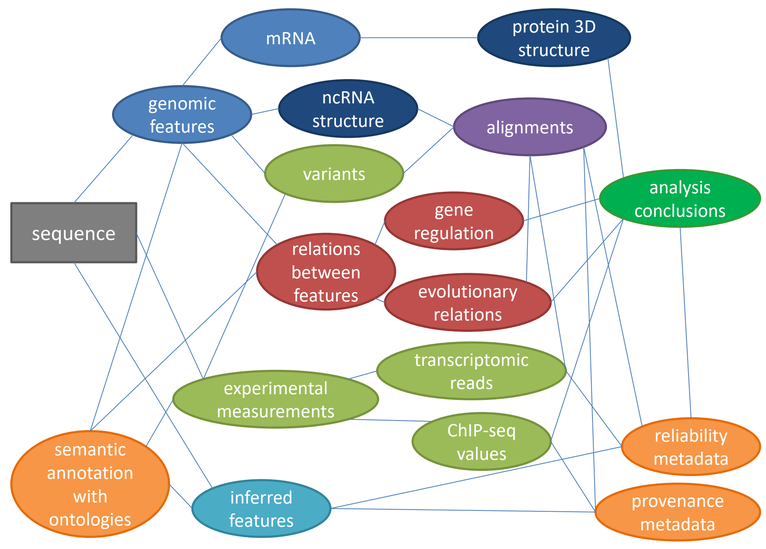
```
{
    "sequence": "MDPLGDTLRRLREAFHAGRTRPAEFRAAQLQGLGRFLQENKQLLHDAL",
    "species": {
        "dbName": "NCBI Taxonomy",
        "accession": "9606",
        "entryUri": "http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606",
        "speciesName": "Human"
    },
    "reference": {
        "dbName": "UniProt",
        "accession": "P43353",
        "entryUri": "http://www.uniprot.org/uniprot/P43353",
        "sequenceVersion": "1",
        "variantAccession": "P43353-1",
        "subsequencePosition": {
            "segment": {
                "min": 1,
                "max": 48
            }
        }
    },
    "name": "Aldehyde dehydrogenase family 3 member B1 (ALDH3B1), N-terminus"
}
```

In BioYAML:

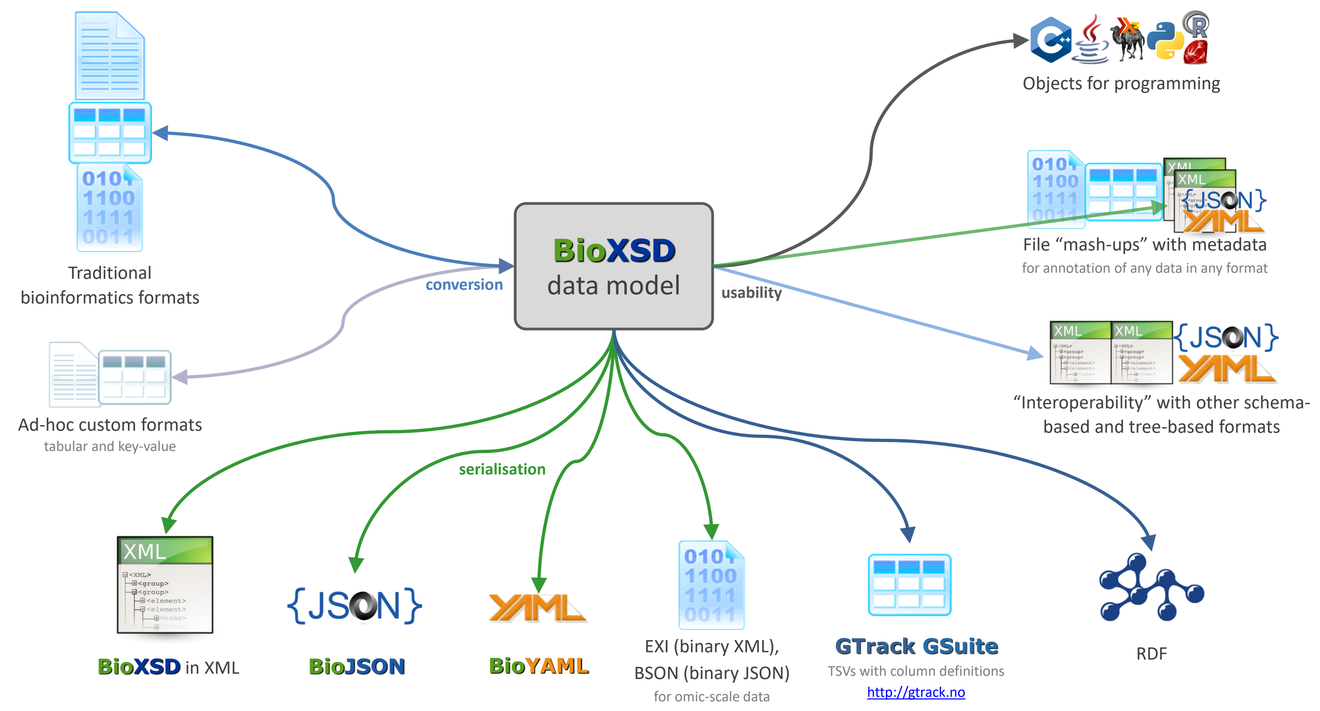
```
---
sequence: MDPLGDTLRRLREAFHAGRTRPAEFRAAQLQGLGRFLQENKQLLHDAL
species:
  dbName: NCBI Taxonomy
  accession: "9606"
  entryUri: http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606
  speciesName: Human
reference:
  dbName: UniProt
  accession: P43353
  entryUri: http://www.uniprot.org/uniprot/P43353
  sequenceVersion: "1"
  variantAccession: P43353-1
  subsequencePosition:
    segment:
      min: 1
      max: 48
name: Aldehyde dehydrogenase family 3 member B1 (ALDH3B1), N-terminus
```

COMPLEX EXAMPLE: BioXSD feature record

BioXSD can represent diverse interconnected features of a sequence, together with related data and metadata, in an integrated data record.



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



BioXSD has been developed as a tree-structured data model and exchange format for basic bioinformatics data, centered around bio-polymer sequence [1,2]. BioXSD allows integration of diverse features, information, measurements, and inferred values about a biological molecule or its part or context, annotated with provenance and reliability metadata, ontology concepts, scientific remarks, and conclusions.

BioJSON and BioYAML are alternatives to the XML serialization of BioXSD, following the same data model. As tree-structured data formats, BioXSD, BioJSON, and BioYAML are particularly suitable for programming in object-oriented languages, and for use with web applications and web APIs (Web services), while at the same time allowing a reasonable level of human readability.

BioXSD|BioJSON|BioYAML are now further developed together with GTrack, GSuite, and BTrack (a family of universal tabular formats for features and metadata [2,3]), under the umbrella of ELIXIR Norway, with an open group of international contributors (http://bioxsd.org/#Contact). The BioXSD|GTrack family of generic, convenient, and interoperable data formats provides an added value for integrative bioinformatics and omics.

[1] Kalaš, M., Puntervoll, P., Joseph, A., Bartaševičiūtė (now Karosiene), E., Töpfer, A., Venkataraman, P., Pettifer, S., Bryne, J.C., Ison, J., Blanchet, C., Rapacki, K. and Jonassen, I. **(2010)**. BioXSD: the common data-exchange format for everyday bioinformatics web services. *Bioinformatics*, **26**, i540-i546. DOI: 10.1093/bioinformatics/btq391 PMID: 20823319 *Open Access*

[2] Gundersen, S., Kalaš, M., Abul, O., Frigessi, A., Hovig, E. and Sandve, G.K. **(2011)**. Identifying elemental genomic track types and representing them uniformly. *BMC Bioinformatics*, **212**, 494. DOI: 10.1186/1471-2105-12-494 PMID: 22208806 *Open Access*

[3] Gundersen S., Kalaš M., Simovski B. *et al.* **(2018)**. The GTrack ecosystem - expressive file formats for analysis of genomic track data [version 1; not peer reviewed]. *F1000Research*, **7**(ELIXIR):270. Poster. DOI: 10.7490/f1000research.1115292.1 *Open Access*

elixir Norwegian Bioinformatics Platform

Oslo University Hospital