XML, XSD, BioXSD

Verena Prade

The Bioinformatics Lab
SS 2013

May 14th, 2013

Extensible Markup Language (XML)

simple text-based format for representing structured information

- both human- and machine-readable
- suitable for Web use
- flexible and extensible
- designed for large scale electronic publishing

XML document

```
1 <?xml version="1.0" encoding="ISO-8859-1"?>
2 <note>
3   <!--This is a comment-->
4   <to>Tove</to>
5   <from>Jani</from>
6   <heading>Reminder</heading>
7   <body>Don't forget me this weekend!</body>
8 </note>
```

- XML declaration
- start and end tags
- one root element
- child and subchild elements
- attributes

Five predefined entities:

< less than "<"
> greater than ">"
& ampersand "&"
' apostrophe '
" quotation mark"

XML document

What are the practical implications of using XML for input and output?

→ Uniform file structure facilitates communication between different tools.

Parsers for XML documents:

Python	ElementTree
Perl	XML::Simple
C++	RapidXml
Java	DOM XML parser

XML documents are extensible

→ adding new tags should pose no problem for parsers

XML Schema

XML Schema

An XML Schema describes the structure of an XML document. The XML Schema language is also referred to as XML Schema Definition (XSD).

An XML Schema defines:

- elements and attributes that can appear in a document
- which elements are child elements
- the order and number of child elements
- whether an element is empty or can include text
- data types for elements and attributes
- default and fixed values for elements and attributes

$\rightarrow$ Allows a validation of the structure of XML documents

What does an XML Schema look like?

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<note xmlns="http://www.w3schools.com"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.w3schools.com note.xsd">
  <!--This is a comment-->
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

What does an XML Schema look like?

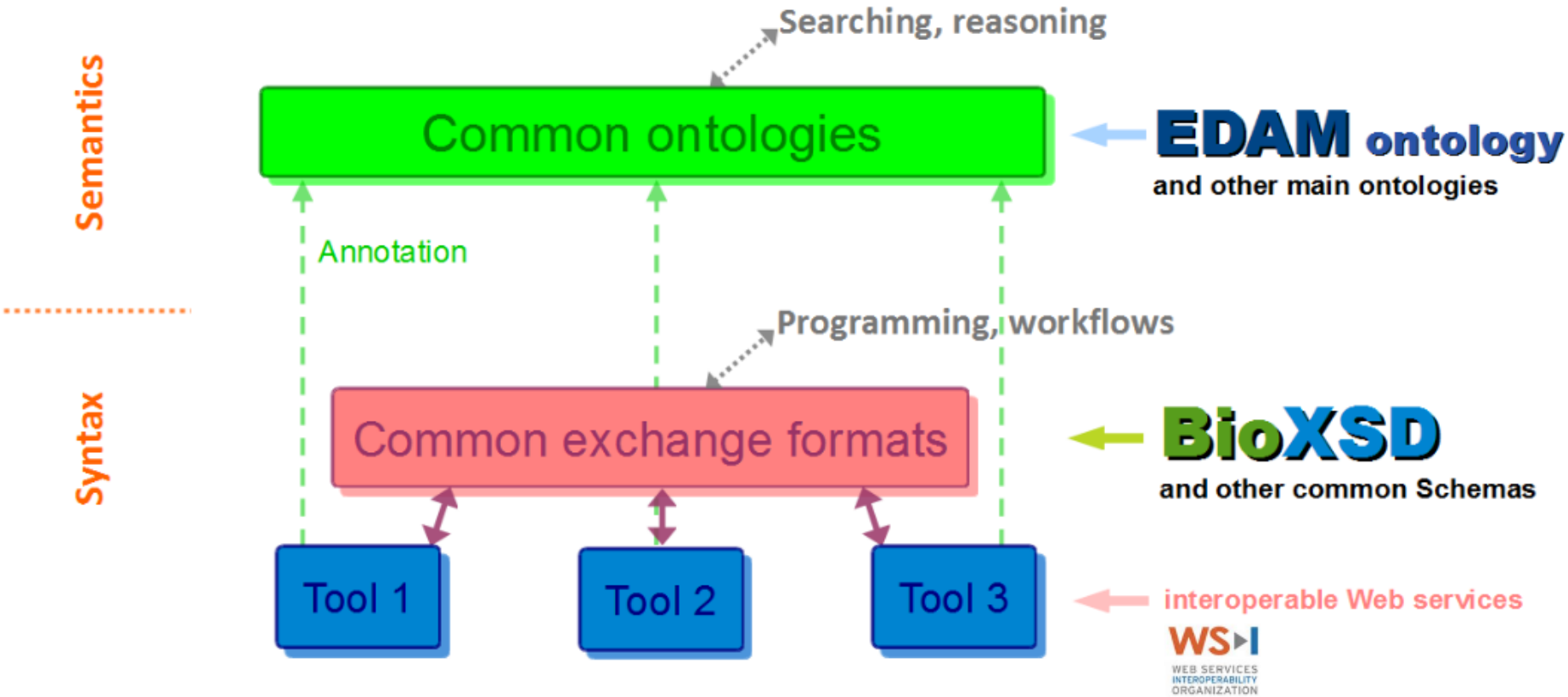
```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<note xmlns="http://www.w3schools.com"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.w3schools.com note.xsd">
  <!--This is a comment-->
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.w3schools.com"
xmlns="http://www.w3schools.com"
elementFormDefault="qualified">

<xs:element name="note">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="to" type="xs:string"/>
      <xs:element name="from" type="xs:string"/>
      <xs:element name="heading" type="xs:string"/>
      <xs:element name="body" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

</xs:schema>
```

What does a document using BioXSD look like?

```
<exampleSequenceRecord>
        <bx:sequence>LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATV
ITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPF
HPYYTIKDFLGLLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALF
LSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAG
XIENY</bx:sequence>
        <bx:species
                speciesName="Elephas maximus maximus"
        />
        <bx:reference
                dbName="GenBank"
                accession="AAD44166"
        />
        <bx:name>Cytochrome b</bx:name>
        <bx:note>Cytochrome b is a subunit of cytochrome bc1, an 11-subunit
 mitochondrial respiratory enzyme</bx:note>
</exampleSequenceRecord>
```

BioXSD and semantic annotations



http://bioxsd.org/bxPoster.pdf

BioXSD schema with referenced ontology

```
<xs:complexType name="AminoacidSequenceRecord" sawsdl:modelReference="http://edamontology.org/data_2047
http://edamontology.org/format_2352">
  <xs:annotation>
    <xs:documentation>Amino-acid sequence record including the unambiguous amino-acid sequence and
optional metadata</xs:documentation>
  </xs:annotation>
  <xs:complexContent>
    <xs:restriction base="GeneralAminoacidSequenceRecord">
      <xs:sequence>
        <xs:element name="sequence" type="AminoacidSequence"/>
        <xs:element name="species" type="Species" minOccurs="0">
          <xs:annotation>
            <xs:documentation>Biological source of the sequence/biopolymer</xs:documentation>
          </xs:annotation>
        </xs:element>
        <xs:element name="reference" type="SequenceReference" minOccurs="0">
          <xs:annotation>
            <xs:documentation>Identification of the sequence record. May refer to data the sequence
originates from: a database entry or an explicit super-sequence.
NB. ID/accession is part of the 'reference'</xs:documentation>
          </xs:annotation>
        </xs:element>
        <xs:element name="name" type="Name" minOccurs="0"/>
        <xs:element name="note" type="Text" minOccurs="0"/>
        <xs:element name="translationData" type="AminoacidTranslationData" minOccurs="0"/>
      </xs:sequence>
    </xs:restriction>
  </xs:complexContent>
</xs:complexType>
```

http://bioxsd.org/technicalDocumentation/BioXSD-1.1/#AminoacidSequenceRecord

Semantic annotation using an ontology - EDAM

Preferred Name *(rdfs:label)*	Sequence record lite (protein)
Definitions *(oboInOwl:hasDefinition)*	A protein sequence and minimal metadata, typically an identifier of the sequence and/or a comment.
ID	data_2047
Full Id	http://edamontology.org/data_2047
Created in	beta12orEarlier
hasDefinition	A protein sequence and minimal metadata, typically an identifier of the sequence and/or a comment.
inSubset	data bioinformatics edam
namespace	data
label	Sequence record lite (protein)
subClassOf	Sequence record (protein) Sequence record lite

http://bioportal.bioontology.org/ontologies/1498?p=terms&conceptid=data_2047

Bibliography

- W3C - Extensible Markup Language (XML).
  http://www.w3.org/
- w3schools.com - XML examples, XML schema.
  http://www.w3schools.com
- XML - Namespaces and XSD Schemas.
  http://www.brainbell.com/tutorials/XML/Namespaces_And_XSD_Schemas.htm
- Kalaš *et al.* (2010).
  BioXSD: the common data-exchange format for everyday bioinformatics web services.
  Bioinformatics (2010) **26** (18): i540-i546.
  http://bioxsd.org/bxPoster.pdf
- EDAM Ontology.
  http://edamontology.org/page