SRA XML Specification

Version 1.2 Draft B Sep 23 2010

National Center for Biotechnology Information – National Library of Medicine

EMBL European Bioinformatics Institute

DNA Databank of Japan

Contents

1 Overview

This document summarizes the proposed changes for Release 1.2 of the Sequence Read Archive (SRA) schemas governing XML metadata. Release 1.2 is an expansion of Release 1.1, which was in introduced in March 2010. The goal of this release is to update choices, introduce new features, and specify a usable Analysis object usable for BAM file submissions. These changes are being introduced with the objective of not invalidating any current valid XML documents.

Major new features in this release are:

- New Analysis schema supports BAM file submissions

1.1 Notice

The features and modalities described in the XML schema DO NOT constitute a statement of features and mechanisms available in the SRA. The schema changes frequently must precede actual implementation. New feature rollouts and functionality changes are made asynchronously with XML schema changes.

1.2 Related Documents

The SRA schema can be obtained from this site:
http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=schema&m=doc&s=schema

1.3 Revision History

2 Explanation of Changes

2.1 Changes to All Documents

2.1.1 LinkType extended

LinkType redefined to include the a choice of the following link types

- SRA_LINK
- URL_LINK
- XREF_LINK
- ENTREZ_LINK
- DDBJ_LINK
- ENA_LINK

2.1.2 New SRA.common.xsd

- SRA.common.xsd factored out of the "COMMON BLOCK" that was included with each SRA schema. New namespace called com: created for commonly used types. [EBI]
- SRA.*.xsd now imports SRA.common.xsd
- Common Block has been refactored to the SRA.common.xsd under namespace com:
- SpotDescriptorType refactored to SRA.common.xsd
- PlatformType refactored to SRA.common.xsd
- ProcessingType refactored to SRA.common.xsd

2.2 Changes to SRA Experiment

2.2.1 Add new instrument models

New instrument values have been added to Experiment :

- "Illumina HiSeq 2000" [Illumina],
- "AB SOLiD System 4" [LifeTech],
- "454 GS Junior" [Roche/454],
- "454 GS FLX Titanium" [Roche/454], to succeed "454 Titanium"

Note that the use of instrument model in Run was deprecated in version 1.1.

2.2.2 Changed EXPERIMENT/PLATFORM/ILLUMINA/CYCLE_COUNT

Changed this to optional field to eliminate need to always specify a deprecated
field. [BI]

2.2.3 Add new library strategy/library selection combinations

New values for LIBRARY_STRATEGY and LIBRARY_SELECTION have been added to Experiment [EDACC]

<LIBRARY_STRATEGY>MRE-Seq</LIBRARY_STRATEGY>
<LIBRARY_SELECTION>Restriction Digest</LIBRARY_SELECTION>

<LIBRARY_STRATEGY>MeDIP-Seq</LIBRARY_STRATEGY>
<LIBRARY_SELECTION>5-methylcytidine antibody</LIBRARY_SELECTION>

<LIBRARY_STRATEGY>RNA-Seq</LIBRARY_STRATEGY>
This is a general choice for sequencing that targets total RNA.

<LIBRARY_STRATEGY>MBD-Seq</LIBRARY_STRATEGY>
<LIBRARY_SELECTION>MBD2 protein methyl-CpG binding domain</LIBRARY_SELECTION>
This combination entails direct sequencing of methylated fractions following enrichment by methyl-CpG binding domain

"WXS" (whole exome sequencing) as a library strategy. [ESP-GO]

2.2.4 Improved documentation for spot descriptor choices.

2.2.5 New Library Source terms

Added TRANSCRIPTOMIC and METAGENOMIC to
EXPERIMENT.LIBRARY_DESCRIPTOR.LIBRARY_SOURCE.
and deprecated EXPERIMENT.LIBRARY_DESCRIPTOR.LIBRARY_SOURCE.NON_GENOMIC.
This is done rather than to offer the choices for biomol, many of which
do not make sense at the level of raw sequencing data (appropriate
for higher level analysis).

2.2.6 PLATFORM nodes

Make all the nodes in PLATFORM consistent to allow for
universal query of PLATFORM/INSTRUMENT_MODEL.
EXPERIMENT.PLATFORM.COMPLETE_GENOMICS.INSTRUMENT_MODEL=none
EXPERIMENT.PLATFORM.PACBIO_SMRT.INSTRUMENT_MODEL=none

2.2.7 Change to sample pool descriptor

SAMPLE_DESCRIPTOR.POOL.MEMBER.READ_LABEL made optional, to support pools that are not barcoded (and therefore don't need a read label). [BI]

2.2.8 Restored expected_number_runs

The attribute expected_number_runs restored (un-deprecated). [EDACC]
This field is actually being used on one roadmap project.

2.2.9 Added TARGETED_LOCUS

Added "TARGETED_LOCUS" as a library element [HMP]

The expectation is that submitters will enter the locus or gene list
for example "16S rRNA V1-V3 region".

2.2.10 Added POOLING_STRATEGY

Added POOLING_STRATEGY as a library element, to help indicate the sample multiplexing intent of the submitter.

2.2.11 Added default_length attribute to SPOT_DESCRIPTOR

Added default_length attribute to EXPECTED_BASECALL and EXPECTED_BASECALL_TABLE so that the parser can assign a default length to a tag when there is no hit in the lookup string or table. [Eugene]

2.2.12 Removed requirement for fields in PROCESSING.QUALITY_SCORES

Removed requirement for deprecated fields in EXPERIMENT.PROCESSING.QUALITY_SCORES [BI]

- <xs:element name="NUMBER_OF_LEVELS" maxOccurs="1" minOccurs="0" type="xs:int"/>
- <xs:element name="MULTIPLIER" maxOccurs="1" minOccurs="0" type="xs:double"/>

2.2.13 New PIPELINE spec in PROCESSING

New element PIPELINE in ProcessingType added to describe the pipeline used in processing the data. This includes a way to specify the sequence of steps in the processing pipeline, programs and their versions, and processing directives.
This was simplified from an earlier proposal, now there is simply a sequence of steps.

2.3 *Changes to Study*

2.3.1 Respecify RELATED_STUDIES

```
<xs:element name="RELATED_STUDIES" minOccurs="0" maxOccurs="1">
   <xs:complexType>
     <xs:sequence>
     <xs:element name="RELATED_STUDY" maxOccurs="unbounded" minOccurs="1">
       <xs:complexType>
         <xs:sequence>
           <xs:element name="RELATED_LINK" type="LinkType" minOccurs="1" maxOccurs="1">
             <xs:annotation>
               <xs:documentation>
                 Related study or project record from a list of supported databases.
                 The study's information is derived from this project record rather
                 than stored as first class information.
               </xs:documentation>
             </xs:annotation>
           </xs:element>
           <xs:element name="IS_PRIMARY" type="xs:boolean" minOccurs="1" maxOccurs="1">
             <xs:annotation>
               <xs:documentation>
                 Whether this study object is designated as the primary source
                 of the study or project information.
               </xs:documentation>
             </xs:annotation>
           </xs:element>
         </xs:sequence>
       </xs:complexType>
     </xs:element>
     </xs:sequence>
   </xs:complexType>
 </xs:element>
```

RELATED_STUDIES is intended to be used as a mechanism to bind the record to the emerging BioProject record (successor to genomeprj record), as well as binding to other resources that track studies (GEO and dbGaP at NCBI, and EGA and ArrayExpress at EBI). [DDBJ]

2.4 *Changes to Sample*

2.5 *Changes to Submission*

In SUBMISSION, added required "schema" attribute to MODIFY action in order to force submitter to specify the namespace of the intended target. "target" is made optional, and will be ignored. Henceforth, the MODIFY source file will contain all the needed references. [BI]

2.6 Changes to Run

2.6.1 Replicated descriptors at Run level

- Replicated SPOT_DESCRIPTOR at the level of Run. If specified at Run, it will override the setting at the level of Experiment.

- Replicated PLATFORM at the level of Run. If specified at Run, it will override the setting at the level of Experiment.

- Replicated PROCESSING at the level of Run. If specified at Run, it will override the setting at the level of Experiment.

2.6.2 New Filetype support

- Added bam as a filetype for RUN.
- Added kar as a supported filetype for RUN, as native SRA format in serialized form.

2.7 Respecified ANALYSIS object

2.7.1 Removed deprecated branches:

- ANALYSIS_TYPE/REPORT
- ANALYSIS_FILES/FILE/filetype/.pdf
- ANALYSIS_FILES/FILE/filetype/.sam (will be delivered in .bam only)

2.8 New SRA Package Object

A new schema SRA.package.xsd has been introduced in order to provide a container for any combination of SRA XML documents, and to allow for applications using SRA objects to aggregate them in any form. SRA packages are not now supported for submission, but eventually will be used in preference to tar archive files.