SRA XML Specification

Version 1.2 Draft K Oct 20 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
- Respecification of processing pipeline and directives
- Reinstantiation of spot descriptor, platform, and processing blocks at the level of SRA Run.
- Addition of choices to many controlled vocabularies

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 for this release can be obtained from this site:
http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA_1-2

1.3 Revision History

Drafts A-K created 17 May 2010 and updated through 20 Oct 2010. Approved for release by INSDC partners 20 Oct 2010. Scheduled for release 01 Nov 2010.

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]

To import this file, do

```
<xs:import schemaLocation="SRA.common.xsd" namespace="SRA.common"/>
```

To use a feature, do

```
<xs:element name="STUDY_ATTRIBUTE" type="com:AttributeType"/>
```

- 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 4 System" [LifeTech],
- "AB SOLiD 4hq System" [LifeTech],
- "AB SOLiD PI System" [LifeTech],
- "454 GS Junior" [Roche/454],
- "454 GS FLX Titanium" [Roche/454], to succeed "454 Titanium"
- "Illumina Genome Analyzer IIx" [Illumina]

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]

- Methylation-Sensitive Restriction Enzyme Sequencing strategy.

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

- Methylated DNA Immunoprecipitation Sequencing strategy.

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

- RNA-Seq strategy

A new choice RNA-Seq was added to LIBRARY_STRATEGY to support the general choice for sequencing that targets total RNA, with the following new choices for LIBRARY_SELECTION (others are possible):

  - CAGE
  - RACE
  - Size fractionation

- Direct sequencing of methylated fractions sequencing strategy.

<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

- Whole exome sequencing strategy

"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 as a way to give further detail to submitters formerly using NON_GENOMIC (which was a holdover choice from the Trace Archive).

2.2.6 PLATFORM nodes

Make all the nodes in PLATFORM consistent to allow for universal query of 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_LOCI block

Added "TARGETED_LOCI" as a library element [HMP, TCGA]. This block allows the submitter to specify one or more gene target(s) or probe set(s) used by the targeted sequencing or hybridization array.
A controlled list will be offered, to consist initially of

- 16S rRNA
- exome
- other

where the submitter can add in free text to identify alternate locus or refine the description..

2.2.10 Added POOLING_STRATEGY

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

- None
- Simple pool
- Multiplexed samples
- Multiplexed libraries
- Spiked library
- Other

This block is added at the level of the library design because in the future sample pools may referenced as an element in BioSamples, rather than a pool spec in SRA experiment.

2.2.11 Added default_ length, base_coord attributes to SPOT_DESCRIPTOR

Added **default_ length**, **base_coord** attributes to EXPECTED_BASECALL and EXPECTED_BASECALL_TABLE. The default_length parameter can specify whether the spot should have a default length for the tag. If provided, the specified number of bases is assigned to this tag regardless of matching criteria. If 0, or not provided, then the tag is "missed" if the match criteria fail. Moreover, submitters should switch to using the EXPECTED_BASECALL_TABLE in preference to EXPECTED_BASECALL. [NCBI, BI]

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. Here is an example of an acceptable processing block under SRA 1.2:

```
<PROCESSING>
  <PIPELINE>
    <PIPE_SECTION section_name="base caller">
      <STEP_INDEX>1.0</STEP_INDEX>
      <PREV_STEP_INDEX>NIL</PREV_STEP_INDEX>
      <PROGRAM>454BaseCaller</PROGRAM>
      <VERSION>1.1.01.20</VERSION>
    </PIPE_SECTION>
    <PIPE_SECTION section_name="SRA conversion">
      <STEP_INDEX>1.1</STEP_INDEX>
      <PREV_STEP_INDEX>1.0</PREV_STEP_INDEX>
      <PROGRAM>toSRA</PROGRAM>
      <VERSION>1.34</VERSION>
    </PIPE_SECTION>
  </PIPELINE>
</PROCESSING>
```

2.2.14 New PROCESSING_DIRECTIVES spec in PROCESSING

A new feature is the explicit enumeration of treatments to the data applied by the submitter, or requested treatments of the data requested by the submitter to be applied by the Archive. Initially this will cover the sample multiplexing directives (no demultiplexing, submitter demultiplexed), but will be expanded in the future to track all treatment requests.

2.3 Changes to Study

2.3.1 New STUDY_TYPE choices

- Exome Sequencing
- Pooled Clone Sequencing.

These were requested by Sanger/EBI.

2.3.2 CENTER_NAME deprecated

The submission and ownership is adequately tracked in the related SUBMISSION object, and the STUDY@center_name attribute.

2.3.3 RELATED_STUDIES

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). This feature should replace the use of PROJECT_ID, a holdover from the Trace Archive, and which has been deprecated.

In order to negotiate a design error in SRA 1.1, a new branch choice has been created called RELATED_STUDY that should be used in preference to the currently effective STUDY, which is now deprecated. The link to a named database was implemented (XRefLinkType) in order to constrain the choice of related project to a named database.

```
<xs:element name="RELATED_STUDY" maxOccurs="unbounded" minOccurs="1">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="RELATED_LINK" type="com:XRefType"
         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>
```

2.4 Changes to Sample

There are no changes to SRA Sample in this revision.

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)

2.7.2 Specified REFERENCE_ALIGNMENT branch

The ANALYSIS/ANALYSIS_TYPE/REFERENCE_ALIGNMENT has been completely specified in order to serve as the metadata container for alignment files delivered in BAM format. Several mechanisms have been furnished to allow submitters to specify the reference sequence. Some additional business rules may also be applied by each Archive to constrain reference choices.

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.

3 Deprecated Fields

SRA 1.2 contains the following fields, branches, and options that should no longer be used in current submissions.

SRA.common.xsd	SPOT_DECODE_METHOD	
SRA.common.xsd	NUMBER_OF_READS_PER_SPOT	
SRA.common.xsd	'454 Titanium'	use '454 GS FLX Titanium'

SRA.common.xsd	'GS 20'	use '454 GS 20'
SRA.common.xsd	'GS FLX'	use 'GS FLX'
SRA.common.xsd	'Solexa 1G Genome Analyzer'	use 'Illumina Genome Analyzer'
SRA.common.xsd	CYCLE_SEQUENCE	use SEQUENCE_LENGTH
SRA.common.xsd	CYCLE_COUNT	use SEQUENCE_LENGTH
SRA.study.xsd	CENTER_NAME	use STUDY@center_name
SRA.study.xsd	PROJECT_ID	use RELATED_STUDIES instead
SRA.study.xsd	RELATED_STUDIES/STUDY	use RELATED_STUDIES/RELATED_STUDY instead
SRA.experiment.xsd	LIBRARY_STRATGEY/BARCODE	use another library strategy
SRA.experiment.xsd	LIBRARY_SOURCE/NON GENOMIC	use METAGENOMIC or TRANSCRIPTOMIC instead
SRA.experiment.xsd	PROCESSING/BASE_CALLS	use PIPELINE instead
SRA.experiment.xsd	PROCESSING/QUALITY_SCORES	use PIPELINE instead
SRA.experimentxsd	@expected_number_spots	
SRA.experimentxsd	@expected_number_reads	
SRA.run.xsd	'_seq.txt, _prb.txt, _sig2.txt, _qhg.txt'	use 'Illumina_native' instead
SRA.run.xsd	@total_spots	
SRA.run.xsd	@total_reads	
SRA.run.xsd	@number_channels	
SRA.run.xsd	@format_code	
SRA.run.xsd	@instrument_model	use PLATFORM/INSTRUMENT_MODEL instead
SRA.run.xsd	@run_file	
SRA.run.xsd	@total_data_blocks	
SRA.submission.xsd	HoldForPeriod	
SRA.submission.xsd	@submission_id	use alias instead
SRA.submission.xsd	@handle	

4 Future Planned Revisions

The next revision is anticipated to be contracting revision (one that potentially invalidates current documents). The main changes will be to remove deprecated fields. This will involve migration of data in anticipation of future schema changes.