SRA XML Specification

Release SRA_1-1 Change Notice

National Center for Biotechnology Information – National Library of Medicine

European Bioinformatics Institute – EMBL

DNA Databank of Japan

Draft L 04 Dec 2009

Contents

1 Overview

This document summarizes the proposed changes for Release 1.1 of the Sequence Read Archive (SRA) schemas governing XML metadata. Release 1.1 is an expansion of Release 1.0, which was in introduced in April 2009. The goal of this release is to patch the XML schema with needed changes while not invalidating current XML implementations.

Major new features in this release are:

- Additional library choices needed for epigenomics
- Additional platform and instrument choices
- Bar code support for pooled and multiplexed samples
- Tightening of specification of run files to allow for improvement of loader programs
- Identification of disused features and options

A second release (SRA 1.2) will be organized with deeper changes that will require migration of existing data and possible changes to client XML generation software is also planned. This release will include a period of public comment and ample time for adjustment and migration.

The third release (SRA 1.3) should then introduce the next round of feature changes.

1.1 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

1.2 Revision History

16 Feb 2009 – Draft A worked out by Rasko Leinonen and Martin Shumway

26 March 2009 – Draft B prepared by Martin Shumway for review by Rasko Leinonen

27 March 2009 – Draft C prepared by Martin Shumway for review by Rasko Leinonen

27 Sep 2009 – Draft F final draft prepared by Martin Shumway for internal review

09 Oct 2009 – Draft G final draft prepared by Martin Shumway for INSDC review

17 Nov 2009 – Draft K draft prepared by Martin Shumway for INSDC review

04 Dec 2009 – Draft L draft prepared by Martin Shumway for INSDC review

1.3 Release Plan

After a period of review existing documents can assume the new schema without change. Following deployment, the SRA capabilities will be built out to take full advantage of the new schema features. Finally, deprecated fields in existing documents will be migrated. The next revision will concentrate on removing deprecated features that have already been migrated from existing documents.

1.4 Glossary

2 Explanation of Changes

2.1 Changes to All Documents

2.1.1 Remove counts from all documents

These fields have proved misleading when bound by the submitter. These fields remain as optional attributes in EXPERIMENT and RUN, but deprecation warnings will be issued for new documents that have these fields bound.

2.1.2 Add to all documents additional link type: XREF_LINK

This is another way to specify external links using the database and accession. This method relies on the archive to construct a proper link, but is less sensitive to changes in the way links are served in the external database.

2.1.3 Add to all documents additional link type: SRA_LINK

This is another way to specify local links visible within the scope of the Home Archive. The intent is that during submission and loading such links will be converted to XREF_LINK where possible.

2.1.4 Add ownership attributes to all documents

The following attributes have been added to all documents: center_name, broker_name.

These allow for establishing a namespace of the center for any document. In addition, refcenter_name has been added to allow for specification of the reference name space.

2.2 Add New Choices to Schema

2.2.1 Add new instrument models

New instrument values have been added to Experiment

- 454 XLR Titanium
- Illumina Genome Analyzer II
- AB SOLiD System 2.0
- AB SOLiD System 3.0

The use of instrument model in Run has been deprecated.

2.2.2 New Study type values

- RNASeq
- Other

2.2.3 Add new library selection values

New values for LIBRARY_SELECTION have been added to Experiment.

- Hybrid Selection
- DNAse

2.2.4 Add new library strategy values

New values for LIBRARY_STRATEGY have been added to Experiment

- Bisulfite-Seq
- DNAse-Hypersensitivity

In addition, BARCODE has been deprecated as it pertains to a pooling strategy and not a library strategy.

2.2.5 Add new library selection values

- Reduced Representation

2.3 Changes to Study

2.3.1 Add RELATED_STUDIES to SRA Study

The SRA study is usually a surrogate object that includes information from one or more studies in other repositories. The SRA_LINK or XREF_LINK mechanisms can be used to call out these dependencies. At some point in the future SRA Study will be migrated over to the new Project repository.

As PROJECT_ID is deprecated, use the RELATED_STUDIES mechanism instead.

2.4 Changes to Sample

2.4.1 Add fields to Sample Name

The following fields have been added to SAMPLE.SAMPLE_NAME in order to create additional ways to unambiguously name a sample:

- SCIENTIFIC_NAME
  Scientific name of sample that distinguishes its taxonomy. Please use a name or synonym that is tracked in the INSDC Taxonomy database. Also, this field can be used to confirm the TAXON_ID setting.

- INDIVIDUAL_NAME - Individual name of the sample. This field can be used to identify the individual identity of a sample where appropriate (this is usually NOT appropriate for human subjects). Example: "Glennie" the platypus.

Documentation for all the sample name fields has been improved, giving better guidance to submitters:

2.4.2 Add Title to Sample

The Sample object now should have a title to make it easier to search. For example: "E. coli K-12 MG1665 genomic sample." Titles need not be unique.

2.4.3 Move Sample Members Table to Experiment

This change means that sample pools will be specified at the level of experiment. Multiplexed sample experiments where each sample is distinguishable by a bar code are listed by sample and bar code. Pooled samples are listed by sample only. Sample can be identified by alias or accession. SAMPLE.MEMBERS has been removed from the schema. A SAMPLE_POOL_DESCRIPTOR has been defined as an option for EXPERIMENT.DESIGN instead of SAMPLE in order to support experiments conducted on sample pools (multiplexed or otherwise).

2.5 Changes to Submission

2.5.1 Submission structure made more flexible

- Submissions may not need FILES section and no longer have to have one even if there are no files.
- Outgoing submission XML may be stripped of CONTACTS, ACTIONS, FILES data because they are not relevant to the user of the Archive.

2.5.2 WITHDRAW to become SUPPRESS

This submission action is actually the GenBank SUPPRESS action.

2.5.3 Add a new action called PROTECT

TO support submission of short read data into protected databases like dbGaP.

2.5.4 Remove HoldUntilPublication

SUBMISSION.ACTIONS.ACTION.HOLD@HoldUntilPublication is to be removed because the feature is underspecified.

2.5.5 Remove CURATE

SUBMISSION.ACTIONS.ACTION.CURATE to be removed, not used.

2.5.6 Remove SUBMISSION.handle

This field is never used.

2.5.7 Remove requestor, request_date

The SUBMISSION.ACTIONS requestor and request_date tags have been deprecated because the submission system will record this info.

2.5.8 Add submission title

SUBMISSION.TITLE would be used in some cases by submitters who are referencing SRA/ERA accession in their publication.

2.5.9 Remove EXCEPTIONS block

SUBMISSION.EXCEPTIONS to be replaced by a dedicated document (SRA.Receipt.xsd) for this purpose.

2.5.10 Rename submission_id to alias

Insert new attribute called SUBMISSION.alias to be consistent with other objects. Deprecate SUBMISSION.submission_id and remove later.

2.5.11 Add links and attributes to Submission

Links and Attributes that are available to all documents are included with Submission. This is to allow for binding of information specific to the submission (but not the content or metadata) to the submission in a flexible way. This information is NOT intended to be used in indexing.

2.6 Changes to Run

The RUN.DATA_BLOCK is made optional in order to redact the submission information from the Run record, in the case where the Run record is displayed to the user of the archive (in Entrez XML, or in the ERA). However, this field continues to be required to process a submission.

The **RUN.run_file** attribute has never been used effectively. It is not needed and the data in it can be dropped from the archive.

2.6.1 Changes to RUN.DATA_BLOCK

Several changes have been made to the DATA_BLOCK itself:

The **DATA_BLOCK** contains loading instructions about the data. It is not needed once the data have been successfully loaded, and does not need to be included in mirrored or downloaded metadata dumps. Therefore, **DATA_BLOCK** has been made technically optional, although it is required on submission.

The new **DATA_BLOCK.serial** attribute will allow for loading of multiple DATA_BLOCKs by indicating the load order. This specification is needed in order for loaders to work with multiple DATA_BLOCK loads.

The new **DATA_BLOCK.FILE.filetype.sra** choice has been added in anticipation of direct submission of sra objects (native SRA archive files).

The new, more specific choices for 454 native file types has been added:
DATA_BLOCK.FILE.filetype.454_native
DATA_BLOCK.FILE.filetype.454_native_seq
DATA_BLOCK.FILE.filetype.454_native_qual

The **DATA_BLOCK.FILE.filetype.Helicos_native** choice has been added to support restricted grammars for Helicos text data.

New options for Illumina native filetypes have been added in order to provide more specificity:
Illumina_native_seq
Illumina_native_prb
Illumina_native_int
Illumina_native_qseq
Illumina_native_fastq
Illumina_native_scarf

More specific choices are offered for SOLiD native file types
DATA_BLOCK.FILE.filetype.SOLiD_native
DATA_BLOCK.FILE.filetype.SOLiD_native _csfasta
DATA_BLOCK.FILE.filetype.SOLiD_native _qual

Finally, a new **DATA_BLOCK.FILE.filetype.tab** file type allows for the specification of per-spot auxiliary sequence data where this is need for loading (for example alternative read_seg settings).

The new **DATA_BLOCK.FILE.READ_LABEL** allows you to associate a given file to named tag(s) in the spot descriptor (for example F1.qseq vs R1.qseq).

The new **DATA_BLOCK.FILE.DATA_SERIES_LABEL** allows you to associate a given file to one or more named data series in the spot descriptor (for example F1.csfasta vs F1.qual). The choices were taken from the SRA Toolkit documentation:

INSDC:read

INSDC:read_filter

INSDC:quality

INSDC:intensity

INSDC:signal

INSDC:noise

INSDC:position

INSDC:clip_quality_left

INSDC:clip_quality_right

INSDC:readname

INSDC:read_seg

Together, **DATA_BLOCK.FILE.READ_LABEL and DATA_BLOCK.FILE.DATA_SERIES_LABEL** should be able to specify most loader configurations we have encountered. Note that these are modeled as elements so that a file can file multiple data series or match multiple read labels.

DATA_BLOCK.FILE.checksum and **DATA_BLOCK.FILE.checksum_method** can be used to specify the checksum of the final component that will be presented to the loader.

The **DATA_BLOCK.FILE.quality_scoring_system** allows the submitter to specify that the incoming data is in log-odds form. This will allow the loader to not have to prescan the file in order to guess which scoring system is being used.

The **DATA_BLOCK.FILE.qualty_encoding** tells whether the quality string is ASCII character, decimal, or hexadecimal encoded. The **DATA_BLOCK.FILE.ascii_offset** allows for the specification of the representation of the basis value (the zero) in the quality data series. Together these parameters can interpret any character based or decimal based quality representation.

The new **DATA_BLOCK.member_name** allows an individual data block among several to be associated with a member of the sample pool. This is being introduced in anticipation of the introduction of sample bar coding support.

2.7 New ANALYSIS object

The ANALYSIS object will contain unstructured submissions of secondary analysis of sequence read objects, including assemblies, alignments, and clean sequence datasets appropriate for submission to dbEST.

2.8 Changes to Experiment

2.8.1 Deprecated SPOT_DECODE_SPEC, NUMBER_OF_READS_PER_SPOT

The SPOT_DECODE_METHOD tag is deprecated because it was underspecified, and because the spot layout needs to be fully specified in every case. The NUMBER_OF_READS_PER_SPOT tag has been deprecated as it is redundant with the READ_SPEC entries.

2.8.2 Decode options added to Spot Descriptor

A new read attribute READ_SPEC.READ_LABEL allows for the naming of tags (F3, R3).

The EXPECTED_BASECALL_TABLE can be used to lookup the combination of tags that can resolve a given spot's relationship with a set of samples in a sample pool.

Added a new READ_SPEC choice branch called RELATIVE_ORDER, which specifies that the read in question is to be found before or after the specified read.

2.8.3 Changes to EXPERIMENT.PLATFORM

For 454, the following fields apply: KEY_SEQUENCE, FLOW_SEQUENCE, FLOW_COUNT.

For Illumina and AB_SOLiD, the following field should be used from now on: SEQUENCE_LENGTH, intended to be the number of bases/colors in the raw sequence (including both mate pairs and any technical reads). CYCLE_SEQUENCE and CYCLE_COUNT are deprecated.

Added placeholder branches for COMPLETE_GENOMICS and PACBIO_SMRT platforms.

Added instrument_model choice for HELICOS (HeliScope).

2.8.4 Changes to EXPERIMENT.PROCESSING

Several unused fields are deprecated: QUALITY_SCORES.NUMBER_OF_LEVELS and QUALITY_SCORES.MULTIPLIER.

2.9 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 Summary of Deprecated Fields

EXPERIMENT.DESIGN.SPOT_DESCRIPTOR.NUMBER_OF_READS_PER_SPOT

EXPERIMENT.DESIGN.SPOT_DESCRIPTOR.SPOT_DECODE_METHOD

EXPERIMENT.expected_number_bases

EXPERIMENT.expected_number_reads

EXPERIMENT.expected_number_spots

EXPERIMENT.PLATFORM.ILLUMINA/AB_SOLID.CYCLE_SEQUENCE, CYCLE_COUNT

EXPERIMENT.PLATFORM.ILLUMINA.instrument_model[Solexa 1G Genome Analyzer]

EXPERIMENT.PLATFORM.LS454.instrument_model[GS 20, GS FLX]

EXPERIMENT.PROCESSING.NUMBER_OF_LEVELS. There should be only one entry for QUALITY_SCORES.

EXPERIMENT.PROCESSING.MULTIPLIER

RUN. DATA_BLOCK.total_spots

RUN.DATA_BLOCK.FILES.FILE.filetype[_seq.txt, _prb.txt, _sig2.txt, _qhg.txt]

RUN.DATA_BLOCK.format_code

RUN.DATA_BLOCK.number_channels

RUN.DATA_BLOCK.total_reads

RUN.instrument_model

RUN.run_file

RUN.total_data_blocks

RUN.total_data_blocks

RUN.total_reads

RUN.total_spots

SAMPLE.members

STUDY.PROJECT_ID

SUBMISSION.submission_id (use alias instead)

SUBMISSION.ACTIONS.ACTION.HOLD.HoldForPeriod

4 Summary of Required Fields

The following fields are optional only at the level of schema, and for the purpose of providing backward compatibility with old documents. New submissions should use these fields.

EXPERIMENT.TITLE – This field is optional in the schema but will eventually be required. As a business rule, new submissions must have this field set to a value.EXPERIMENT.PLATFORM.ILLUMINA.SEQUENCE_LENGTH – This field is optional in the schema but will be required for new submissions.

SAMPLE.TITLE – This field is optional in the schema but will eventually be required. As a business rule, new submissions must have this field set to a value.

SUBMISSION.CONTACTS – This is optional in the schema and may not be reproduced by the Archive because of the private nature of the content. However, on submission the Archive will require a CONTACTS section.

SUBMISSION.ACTIONS – This is optional in the schema and may not be reproduced by the Archive because of irrelevance. However, on submission the Archive will require a ACTIONS section.

SUBMISSION.ACTIONS.ACTION.HOLD – You must specify a date using the HoldUntilDate attribute.

STUDY.RELATED_STUDIES– This is optional in the schema but required in order to identify the source of the study record information.

RUN.DATA_BLOCK – This is optional in the schema but required for a submission containing run data to be processed.

5 Summary of Impending Changes

5.1 Impending Changes to SUBMISSION – SRA.submission.xsd

Changes expected in next major release

Remove HoldUntilPublication option
Remove submission_id, use alias instead
Remove deprecated fields

5.2 Impending Changes in SAMPLE – SRA.sample.xsd

Changes expected in next major release

Require SAMPLE.TITLE
Require SCIENTIFIC_NAME or TAXON_ID

5.3 Impending Changes in RUN – SRA.run.xsd

Changes expected in next major release

Remove deprecated fields

5.4 Impending Changes in EXPERIMENT – SRA.experiment.xsd

Changes expected in next major release

Require EXPERIMENT.TITLE
Allow only one copy of QUALITY_SCORES
Require ILLUMINA.SEQUENCE_LENGTH for Illumina platform choice

5.5 Other Changes

Add support for the CompleteGenomics, PacificBioSciences sequencing platforms.

6 Summary of Future Changes

Future work will address the following issues:

- RUN.SPOT_DESCRIPTOR specialization to allow for better aggregation of runs to libraries.
- EXPERIMENT.PLATFORM respecification
- EXPERIMENT.PROCESSING respecification
- EXPERIMENT links specification to allow for relationships between experiments.