SRA XML Specification Version 1.0 (Deprecated)

Created: April 9, 2009; Updated: March 10, 2010.

<table>
<thead>
<tr>
<th>Status</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Date</td>
<td>2009-04-09</td>
</tr>
<tr>
<td>Inactive Date</td>
<td>2009-11-26</td>
</tr>
<tr>
<td>Scope</td>
<td>INSDC SRA</td>
</tr>
</tbody>
</table>

1 Overview

This document summarizes the proposed changes for Release 1.0 of the Short Read Archive (SRA) schemas governing XML metadata. Release 1.0 is a slight change over Release 0.8, which was introduced in June 2008. The goal of this release is to patch the XML schema with needed changes while not invalidating current XML implementations.

A second release 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.

1.1 Related Documents

1.2 Revision History

16 Feb 2009 - 27 Mar 2009 – Drafts A, B, C worked out by NCBI, EBI, DDBJ

2 Changes

2.1 Remove expected counts from Experiment and Run

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.

2.2 Add center_name attribute to Experiment and Sample

This allows the submitter to establish ownership of the record and allows the Archive to correctly ascribe ownership.
2.3 Add new instrument models

New instrument values have been added to Experiment and Run:

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

2.4 New Study type values

- RNASeq
- Other

2.5 Add new library selection values

New values for LIBRARY_SELECTION have been added to Experiment.

- Hybrid Selection
- DNAse

2.6 Add new library strategy values

New values for LIBRARY_STRATEGY have been added to Experiment

- Bisulfite-Seq
- DNAse-Hypersensitivity

2.7 Add new library selection values

- Reduced Representation

2.8 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.

2.9 Drop run_file from RUN

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.

2.10 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
2.11 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.

2.12 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. The optional DEFAULT_MEMBER identifies the sample to use when none of the specified bar codes matches the reads, usually due to sequencing error.

2.13 WITHDRAW to become SUPPRESS
This submission action is actually the GenBank SUPPRESS action.

2.14 Add a new action called PROTECT
TO support submission of short read data into protected databases like dbGaP.

2.15 Remove HoldUntilPublication
SUBMISSION.ACTIONS.ACTION.HOLD/HoldUntilPublication is to be removed. This is too hard to implement.

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

2.17 Remove SUBMISSION.handle
Not used.

2.18 Remove CLOSE
Not used.

2.19 Remove requestor, request_date
SUBMISSION.ACTIONS.requestor, SUBMISSION.ACTIONS.request_date, Not needed as the submission system will record these.

2.20 Remove handle
SUBMISSION.handle is not used.
2.21 Add submission title

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

2.22 Add broker_name

Identity of broker authority responsible for the submission. This is used to determine ownership and editorial authority, hold and release control, and future editing capability.

2.23 Remove EXCEPTIONS block

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

2.24 Rename submission_id to alias

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

2.25 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.

2.26 Master Study for study

This is to encode the situation where the information of the study is derived from another record, and the values for the study record are derived from the master. Possible values are:

None – (if there is no master study)
INSDC – Genome Project id
GEO – Gene Expression Omnibus at NCBI
ArrayExpress – Array Express resource at EBI
dbGaP – Genotype and Phenotype resource at NCBI
Bioinvestigation Index – BioInvestigation resource at EBI

2.27 RUN_DATA_BLOCK

This section 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.

2.28 RUN.DATA_BLOCK.ADDRESS

This is a new section to describe the address of the run data provided with the submission, when (a) the portion to be loaded is a subset of the data provided, or (b) the data are provided in uncontained format (fastq or native) and the address of the reads needs to be communicated to the loader. The default behavior when this spec is not provided is to load all data encountered in the submitted run data file.

The previous tags that tracked this information, DATA_BLOCK name, sector, region, have been deprecated. The new scheme is specific to the platform but also can be queried in a database.

2.29 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.

2.30 RUN.SPOT_DESCRIPTOR.SPOT_DECODE_SPEC

This tag was removed because it was underspecified, and because the spot layout needs to be fully specified in every case.

2.31 SPOT_DESCRIPTOR moved to Run

The important section describing the layout of the spot sequence has been moved to Run for a variety of reasons:

- In some cases the layout is only known at run time
- In some cases the layout must be changed in response to QC
- In some cases the layout may be specific to a group of reads that define the run

Submissions containing EXPERIMENT.SPOT_DESCRIPTOR will continue to be processed for a time, but new submissions should switch to the Run based location of this block.

2.32 New section RUN.PLATFORM

The Run object now has a PLATFORM block, which includes information from the old EXPERIMENT.PLATFORM as well as new settings such as RAW_SEQUENCE_LENGTH for Illumina and SOLiD in order to indicate the number of bases for these fixed length platforms. New submitters should switch platform details from Experiment to Run. EXPERIMENT.PLATFORM will remain, but in a cut down version, as the highest level indicator of what sequencing platform the experiment is targeting.
The reasons for this change include:

- The instrument model needs to be defined in one place only
- Parameters such as number of flows or bases are run specific and only known at run time.
- Parameters such as flow sequence and color matrix may be static, but are closely related to the interpretation of the run data and therefore should be located there.

This change comes at the cost of redundancy with respect to the runs in the experiment. However, the experiment can be presented as the union of the platform settings, which should be consistent.

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

For Illumina, the following fields apply: SEQUENCE_LENGTH, intended to be the number of bases in the raw sequence (including both mate pairs and any technical reads).

2.33 New section RUN.PROCESSING

The Run object now has a PROCESSING block, which includes information about what pipeline was used to process the run data. As this is not always known until run time, the data should be bound at the level of run. Precise details of the processing pipeline can be embedded in the run container file (SFF or SRF). Some aspects of the former EXPERIMENT.PROCESSING have been refactored into the RUN.PLATFORM block, such as RUN.SEQUENCE_SPACE and RUN.PLATFORM.QUALITY_MODEL.

The design for specifying the run processing pipeline is flexible, and allows for multiple pipe sections arrayed in a workflow. The workflow commands allow for any procedural workflow to be specified. The optional COMMAND tag can be used to specify a complete workflow (with branches and loops), otherwise the process is executed in the order specified by the PROCESS_INDEX tag.

Normally, a single entry will be sufficient, as the processing pipeline is the manufacturer's standard protocol.

3 Summary of Deprecated Fields

EXPERIMENT.expected_number_spots
EXPERIMENT.expected_number_reads
EXPERIMENT.expected_number_bases
RUN.DATA_BLOCKS.DATA_BLOCK.total_spots
RUN.DATA_BLOCKS.DATA_BLOCK.total_reads
RUN.instrument_model[Solexa 1G Genome Analyzer]
RUN.instrument_model[Solexa 1G Genome Analyzer]
EXPERIMENT.PLATFORM.LS454.instrument_model[GS 20]
EXPERIMENT.PLATFORM.LS454.instrument_model[GS FLX]
RUN.run_file

STUDY.PROJECT
RUN.total_spots
RUN.total_reads
RUN.number_channels
RUN.format_code
RUN.total_data_blocks
RUN.run_file
SAMPLE.members
DATA_BLOCK.name
DATA_BLOCK.sector
DATA_BLOCK.region

4 Summary of Required 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.

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.

STUDY.MASTER_STUDY – This is optional in the schema but required in order to identify the source of the study record information, or None if there is no master record.

RUN.DATA_BLOCK – This is optional in the schema but required for a submission to be processed.
RUN.DATA_BLOCK.ADDRESS – This spec is required when the data are submitted in an uncontained format (such as fastq or a native format).