Use this form to tell us important information about this document, then start the text on the following page. All information you give in this form will appear in the document.

Document Information

Document Title	Created (YYYY-MM-DD)	Updated (YYYY-MM-
SRA XML Schema 1.5 Release Notes	2012-06-05	2012-06-07

Author Information

Given Name(s)	Last Name	Suffix	Degrees	Affiliation	Email

Use one row for each author. List authors in order of appearance in the document. Add rows to add more authors.

SRA XML Schema 1.5 Release Notes

Draft B – 07 Jun 2012

Status	Active
Active Date	2012-09-04
Inactive Date	
Scope	INSDC SRA

Table Of Contents

Overview

This document summarizes the proposed changes for Release 1.5 of the Sequence Read Archive (SRA) schemas governing XML metadata. This schema will be used by the SRA archive instances and has been developed under the auspices of the International Nucleotide Sequence Database Collaboration (INSDC, insdc.org).

Although there are a few expansionary elements, Release 1.5 is mainly a contractionary change over Release 1.4, which was in introduced in May 2012. These changes are being introduced with the intention of invalidating previous XML schema releases.

Major new features in this release are:

- Add new file types to support the CompleteGenomics platform
- Require LibraryStrategy element
- Rationalize some Library choices
- Remove many deprecated elements and structures
- Introduce the SRA.Reference.xsd, which specifies reference sequences used in aligned SRA objects.

Notice

The features described in the SRA 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. Each SRA implementation by INSDC partners may impose additional business rules not reflected in the schema.

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_1-5b

Revision History

Draft B- 2012-06-07 submitted for approval by INSDC partners

Explanation of Changes

New ReferenceType

A new schema document has been introduced called SRA.Reference.xsd, which encodes the ReferenceType. This data structure defines the various ways to represent reference sequences.

Changes to LibraryDescriptorType

Add “ncRNA-Seq” library strategy for non-coding RNA-seq

A new library strategy called ncRNA-Seq has been introduced for non-coding RNA-Seq other than miRNA-Seq.

"Discovery of other non-coding RNA types, including post-translation modification (snRNA(small nuclear RNA) or snoRNA (small nucleolar RNA)) and expression regulation (siRNA(small interfering RNA) or piRNA/piwiRNA(piwi-interacting RNA))."

Remove Certain Library Selections

The choices for Cot-filtration have been consolidated into “repeat fractionation” using “Cot filtration or other means”. This change remaps CF-M, CF-H, CF-T, CF-S to “repeat fractionation”.

Remove Certain Library Source

NON-GENOMIC selection has been removed, to be replaced by TRANSCRIPTOMIC or METATRANSCRIPTOMIC.

Require LibraryStrategy

LibraryStrategy is now required. The value Unspecified can be used to migrate the existing records that do not have this element.

Change to RunType

In order to support support richer submissions from CompleteGenomics, the filetypes table has been augmented with these new filetypes:

CompleteGenomics_native_MAP
CompleteGenomics_native_ASM
CompleteGenomics_native_LIB
CompleteGenomics_merged_BAM
CompleteGenomics_evidence_BAM
CompleteGenomics_fastq

The following fields have been removed:

Remove already deprecated 'instrument_model' attribute.

Remove already deprecated 'run_file' and 'total_data_blocks' attributes.

Remove 'total_spots', 'total_reads', 'number_channels', 'format_code' deprecated attributes form DATA_BLOCK element.

Changes to SpotDescriptorType

The following elements have been eliminated:

SpotDescriptorType: Remove READ_SPEC/CYCLE_COORD.

SpotDescriptorType: Remove EXPECTED_BASECALL element (use EXPECTED_BASECALL_TABLE instead).

SpotDescriptorType: Remove ADAPTER_SPEC element.

SpotDescriptorType: Remove already deprecated 'SPOT_DECODE_METHOD' element.

SpotDescriptorType: Remove already deprecated 'NUMBER_OF_READS_PER_SPOT' element.

SpotDescriptorType: Remove already deprecated 'CYCLE_COORD' element.

SpotDescriptorType: Remove deprecated EXPECTED_BASECALL element.

Changes to PlatformType

The following elements have been eliminated:

PlatformType: Remove optional SEQUENCE_LENGTH from ILLUMINA and ABI_SOLID platforms.

PlatformType: Remove optional FLOW_COUNT from LS454, HELICOS platforms.

PlatformType: Remove deprecated CYCLE_COUNT element from ILLUMINA and ABI_SOLID platforms.

PlatformType: Remove optional CYCLE_SEQUENCE element from ILLUMINA platform.

PlatformType: Remove optional KEY_SEQUENCE and FLOW_SEQUENCE element from LS454 platform and FLOW_SEQUENCE from HELICOS platform.

PlatformType: Remove optional COLOR_MATRIX and COLOR_MATRIX_CODE from ABI_SOLID platform.

Change to ExperimentType

The following elements have been eliminated:

Remove optional attribute 'expected_number_runs'.

Remove already deprecated attributes 'expected_number_spots' and 'expected_number_reads'.

Remove PROCESSING/BASE_CALLS element.

Remove PROCESSING/QUALITY_SCORES element.

Remove deprecated instrument models for LifeTech 5500 series instruments in favor of correct 5500 series instrument names.

Changes to SubmissionType

The following fields have been changed:

remove SUBMISSION/FILES

Remove 'target' attribute in MODIFY action.

Remove 'CLOSE' action.

Remove 'notes' attribute from ADD, MODIFY, SUPPRESS, HOLD, RELEASE, VALIDATE actions.

remove SUBMISSION/FILES

Make 'schema' attribute mandatory for MODIFY action.

Remove 'HoldForPeriod' attribute from HOLD action.

Deprecated Fields

SRA 1.5 contains the following fields, branches, and options that should no longer be used in current submissions.

Field	Notes

Notes

1. No deprecated fields in this version.

Future Planned Revisions

The next revision, SRA 2.0, will be a change of the SRA data model in order to accommodate newer sequencing and alignment technologies.

Figures, Tables and Boxes Appendix (do not delete)

Place numbered figures, tables and boxes (referred to from the main text) below.

“In-line” figures (e.g. equations) and tables should be placed within the main text in their desired final location.

Boxes can have a single level of sections; the titles for these sections should be marked up in “Box subhead” style.