Using the SRA Identifier Block

30 Apr 2012 Draft E

Overview

The purpose of the SRA Identifier block is to capture in one place all keys that are used as IDs. An ID can identify exactly one record within a context. A record may have multiple IDs. A record's ID must be unique within a context, and all objects in a context must have an ID. These properties do not hold for "names" or other monikers.

The SRA Identifier block contains the following identifiers:

- PRIMARY_ID – Primary key to an INSDC database.
- SECONDARY_ID – Secondary key or defunct primary key to an INSDC database
- EXTERNAL_ID – Identifier from another database qualified by a namespace.
- SUBMITTER_ID – Local identifier qualified by a namespace.
- UUID – Universally unique identifier which requires no namespace.

The identifiers block contains a set of identifiers. The identifiers may occur in any order or combination so long as exactly one PRIMARY_ID is present.

Goals

- Consolidate use of identifiers for each SRA document
- Distinguish between accessions and named IDs
- Add support for UUIDs
- Improve flexibility for submitter identification of records

Features

- Tracks archived assigned ids
- Tracks submitter assigned ids
- Tracks 3rd party assigned ids including catalog ids
- Can support secondary ids and aliases
- Can support UUIDs
- Tracks alternate or secondary ids assigned by different archives
- Tracks replacement of previously active records
- Tracks equivalence between records

Design

The IdentifiersType is defined in the SRA.common.xsd schema, please look in the following location(s):

- file:///home/shumwaym/proj/SRA/sra/doc/SRA.common.xsd
- https://svn.ncbi.nlm.nih.gov/viewvc/toolkit/trunk/internal/trace_archives/sra/doc/SRA.common.xsd?view=markup
- http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA_1-4e/SRA.common.xsd?view=co
- http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/sra/doc/SRA_1-4/SRA.common.xsd?view=co

- Here is the relevant type code from SRA.common.xsd:

```
<xsd:complexType name="IdentifierNodeType">
    <xsd:simpleContent>
        <xsd:extension base="xs:string">
            <xsd:attribute name="label" use="optional" type="xs:string"/>
        </xsd:extension>
    </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="AccessionType">
    <xsd:simpleContent>
        <xsd:extension base="com:IdentifierNodeType"/>
    </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="NameType">
    <xsd:simpleContent>
        <xsd:extension base="com:IdentifierNodeType">
            <xsd:attribute name="namespace" use="required" type="xs:string"/>
        </xsd:extension>
    </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="UUIDType">
    <xsd:simpleContent>
        <xsd:extension base="com:IdentifierNodeType"/>
    </xsd:simpleContent>
</xsd:complexType>

<xsd:complexType name="IdentifierType">
    <xsd:sequence>
        <xsd:element name="PRIMARY_ID" type="com:AccessionType" minOccurs="1" maxOccurs="1"/>
        <xsd:element name="SECONDARY_ID" type="com:AccessionType" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element name="EXTERNAL_ID" type="com:NameType" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element name="SUBMITTER_ID" type="com:NameType" minOccurs="0" maxOccurs="unbounded"/>
        <xsd:element name="UUID" type="com:UUIDType" minOccurs="0" maxOccurs="1"/>
    </xsd:sequence>
</xsd:complexType>
```

Data Types

The IdentifierNodeType abstract type extends xs:string with the following attributes:

- **label** – whether and how to display a tag string.

Four concrete types subclass IdentiferType in order to suggest the business use of the identifier:

AccessionType – A key in an INSDC primary database.

NameType –A key in an external database.

The following attributes are required:

- **namespace** – Namespace (database) of the external name

UUIDType – A key that is universally unique and requires no namespace.

Data Structure

PRIMARY_ID – A primary identifier, or key, in the INSDC primary database (accession). Example: SRR000123. Exactly one primary identifier is required in every IDENTIFIER block. This value is equivalent to the document/@accession attribute.

SECONDARY_ID – A foreign key in the INSDC primary database (accession), or a defunct primary key in the INSDC primary database. Example: PRJNA41443. Any number of secondary identifiers may be present.

EXTERNAL_ID – A key in an external database qualified by the name of the database. Example: Coriell NA12878. Any number of external names may be present.

SUBMITTER_ID – A key that resolves within the current set of documents. Exactly one local name must be present on submission. Local names are not needed for data download or exchange between archives. This value is equivalent to the (document/@alias, document/@center_name) attribute tuple.

UUID – A key that is universally unique and needs no namespace. UUIDs are not used by the Archive but rather are provided as part of the SRA xml schema to serve downstream applications, including non-INSDC SRA mirrors.

Compatibility

It is intended that the existing NameGroup and RefNameGroup types will continue to remain in use for backwards compatibility.

Semantics

The IdentifierType is implemented by each SRA archive with additional business rules governing use of namespaces and scope of identifiers.

Controlled Namespaces

Most namespaces are not interpreted and only apply to the current submission. The exceptions are listed here:

Reserved Namespaces (ID class)	**Type**	**use**
ega, arrayexpress	EBI ENA database	Identify federated resource
coriell, atcc	sample vendor	Identify externally curated samples
TCGA, TARGET, EMMES_HMP, …	Sample namespaces	Allow namespace/samplename identifiers
BI, BCM, BCCAGSC, WUGSC, JGI, …	Center namespaces	Identify records within a center's namespace See the current centers list at : ftp://ftp.ncbi.nlm.nih.gov/sra/reports/Centers/centers.tab

Replacement tracking

The IdentifierType can be used to name record(s) replaced (taken over) by the current record. The transitive closure of these replacing relations is a set of currently active records with replaced descendants. The converse relation (replaced by) can be computed from this forest so it is not tracked explicitly.

Migration tracking

The replaced by relation would be tracked in the case where the record was replaced by a record in a new database (migrated), for example biosample or bioproject. Another case is where a record was moved from one INSDC SRA to another and thereby received a replacement accession.

Persistence

One goal of the IDENTIFIERS block is to document data migration, replacement, and equivalency relationships independently of the life cycle of the record, so that Archive users who form dependencies on a certain SRA record can always recover the relationship to other records even if the record has been suppressed.

Use Cases

Data Migration

The IDENTIFIERS block can be used to manage the transition of metadata from one record to another and provide a trackback mechanism to recover previous incarnations. This would include:

- Tracking a record in the archive (or prior to archiving) with a submitter supplied identifier.
- Tracking a record's identifier before and after a data migration.
- Tracking a record's identifier before and after a data consolidation.
- Tracking a changes in an identifier used for a dependency

Data Replacement

The IDENTIFIERS block can be used to indicate that the content has been replaced, and identify the previous record that represented the content. A run may have been mis-loaded due to errors in the original load process or a misrepresentation of the metadata that caused the data to be interpreted differently. If the result of the mis-load is an SRA archive image that is substantially different then the run's accession will be replaced. Another example is where duplicate runs have been discovered, and each run can be mapped to its duplicates although only one of them is retained in the archive.

Data Equivalency

The IDENTIFIERS block can be used to point to records that are equivalent and can be used interchangeably. An example is the BioProject and SRA study identifiers, which for a time will both be active identifiers of a study record (until migration from SRA study to BioProject is completed). Another example is where equivalent records have been discovered in multiple SRA instances. This would happen when a submitter has sent the same submission to both NCBI and EBI, for example. Over time, the INSDC

may elect to retain one instance and suppress the other one, but the ID block can be used to maintain the equivalence relation.

Examples

SRA document identifiers

The document can contain IDENTIFIERS block in co-existence with existing NameGroup attribute group :

```
<RUN xmlnamespace="" run_center="BI" run_date="2011-08-04T04:00:00Z" instrument_name="SL-HAC">
    <IDENTIFIERS>
      <PRIMARY_ID>SRR354028</PRIMARY_ID>
      <SUBMITTER_ID  namespace="BI" >BI.PE.110804_SL-HAC_0370_BFCB02H8ACXX.6.UNMATCHED.srf</SUBMITTER_ID>
    </IDENTIFIERS>
```

The document can contain IDENTIFIERS block in lieu of existing NameGroup attribute group:

```
<RUN>
    <IDENTIFIERS>
      <SUBMITTER_ID  namespace="BI" >BI.PE.110804_SL-HAC_0370_BFCB02H8ACXX.6.UNMATCHED.srf</SUBMITTER_ID>
      <PRIMARY_ID>SRR354028</PRIMARY_ID>
    </IDENTIFIERS>
```

This gives a migration path for adoption of Identifier block in place of the name group attributes group, or a method for reverse construction of the NameGroup attributes from the ID block.

SRA study reference

Document dependency references to other documents can be encoded with or without the NameGroup attributes.

```
    <STUDY_REF accession="SRP009022" refcenter="BI" refname="Ceratotherium_simum_simum_WGS">
          <IDENTIFIERS>
            <PRIMARY_ID>SRP009022</PRIMARY_ID>
            <SECONDARY_ID>PRJNA74583</SECONDARY_ID>
            <SUBMITTER_ID namespace="BI" >Ceratotherium_simum_simum_WGS</SUBMITTER_ID>
          </IDENTIFIERS>
    </STUDY_REF>

    <STUDY_REF>
          <IDENTIFIERS>
            <PRIMARY_ID>SRP009022</PRIMARY_ID>
            <SECONDARY_ID>PRJNA74583</SECONDARY_ID>
            <SUBMITTER_ID namespace="BI" >Ceratotherium_simum_simum_WGS</SUBMITTER_ID>
          </IDENTIFIERS>
    </STUDY_REF>
```

SRA Sample Reference

Where the sample reference is a simple reference, this can be represented with the Identifiers block:

```
        <SAMPLE_DESCRIPTOR>
          <IDENTIFIERS>
            <PRIMARY_ID>SRS293911</PRIMARY_ID>
            <SUBMITTER_ID namespace="JGI">10908</SUBMITTER_ID>
          </IDENTIFIERS>
        </SAMPLE_DESCRIPTOR>
```

Where the sample is a pool and each member is identified with a sample id an Identifiers block can be used instead of (or in addition to) the NameGroup attribute group. This eliminates the need to always provide a default sample where the sample is not part of the pool (typically unidentified organism). In this example the data are partitioned between a list of MEMBERs with no leftovers.

```
<SAMPLE_DESCRIPTOR>
    <POOL>
        <MEMBER proportion="1" member_name="tagged_908_TGCTCGAC">
            <READ_LABEL>barcode</READ_LABEL>
            <IDENTIFIERS>
              <PRIMARY_ID >SRS267431</PRIMARY_ID>
              <SECONDARY_ID >SAMN739917</SECONDARY_ID>
              <SUBMITTER_ID namespace="BI" >478560.5885.New Tech Library.SDZICR_KB13650</SUBMITTER_ID>
            </IDENTIFIERS>
        </MEMBER>
        . . .
    </POOL>
</SAMPLE_DESCRIPTOR>
```

In this example, there exists a DEFAULT_MEMBER, which can hold the reference to the sample for those reads that cannot be assigned to a specific sample.

```
<SAMPLE_DESCRIPTOR>
     <POOL>
         <MEMBER member_name="M10_V2">
             <READ_LABEL read_group_tag="M10">barcode</READ_LABEL>
             <READ_LABEL read_group_tag="V2">rRNA_primer</READ_LABEL>
             <IDENTIFIERS>
               <PRIMARY_ID >SRS008987</PRIMARY_ID>
               <SECONDARY_ID >SAMN 6821</SECONDARY_ID>
               <SUBMITTER_ID namespace="CCME" > M10</SUBMITTER_ID>
             </IDENTIFIERS>
         </MEMBER>
           . . .
         <DEFAULT_MEMBER member_name="">
             <IDENTIFIERS>
               <PRIMARY_ID >SRS0001216</PRIMARY_ID>
               <SECONDARY_ID >SAMN 6214</SECONDARY_ID>
               <SUBMITTER_ID namespace="CCME" >fierer_hand_study_default</SUBMITTER_ID>
             </IDENTIFIERS>
         </DEFAULT_MEMBER>
      </POOL>
</SAMPLE_DESCRIPTOR>
```

BioSample Reference

The successor BioSample record can be identified alongside the SRA sample accession, as in:

```
<SAMPLE_DESCRIPTOR>
  <IDENTIFIERS>
     <PRIMARY_ID >SRS267431</PRIMARY_ID>
     <SECONDARY_ID >SAMN739917</SECONDARY_ID>
     <SUBMITTER_ID namespace="BI">478560.5885.New Tech Library.SDZICR_KB13650</SUBMITTER_ID>
   </IDENTIFIERS>
```

When the SRA Sample record becomes secondary, the PRIMARY_ID/SECONDARY_ID identifiers can indicate this:

```
<SAMPLE_DESCRIPTOR>
  <IDENTIFIERS>
     <PRIMARY_ID >SAMN739917</PRIMARY_ID>
     <SECONDARY_ID >SRS267431</SECONDARY_ID>
     <SUBMITTER_ID namespace="BI">478560.5885.New Tech Library.SDZICR_KB13650</SUBMITTER_ID>
   </IDENTIFIERS>
```

BioProject Reference

The successor BioProject record can be identified alongside the SRA Study accession, as in:

```
<STUDY_REF>
      <IDENTIFIERS>
```

```
            <PRIMARY_ID >SRP010976</PRIMARY_ID>
            <SECONDARY_ID >PRJNA74601</SECONDARY_ID>
            <SUBMITTER_ID namespace="JGI">10909</SUBMITTER_ID>
          </IDENTIFIERS>
    </STUDY_REF>
```

When the SRA Study record becomes secondary, the IDENTIFIERS block can reflect this:

```
    <STUDY_REF>
          <IDENTIFIERS>
            <PRIMARY_ID>PRJNA74601</PRIMARY_ID>
            <SECONDARY_ID>SRP010976</SECONDARY_ID>
            <SUBMITTER_ID namespace="JGI">10909</SUBMITTER_ID>
          </IDENTIFIERS>
    </STUDY_REF>
```

When the SRA Study record becomes defunct this fact will be reflected in the SRA database and livelist (not in the IDENTIFIER block).

Replaced Record

The information that a certain record has been replaced is not indicated in the IDENTIFIERS block, but is tracked in the SRA database and livelist.

```
<RUN run_date="2008-11-24T23:08:44Z" instrument_name="GA-5">
    <IDENTIFIERS>
      <PRIMARY_ID>SRR292241</PRIMARY_ID>
    </IDENTIFIERS>
```

Replacer Record

This example shows how a record, SRR390728, replaces a predecessor SRR292241:

```
<RUN run_date="2008-11-24T23:08:44Z" instrument_name="GA-5">
    <IDENTIFIERS>
      <PRIMARY_ID>SRR390728</PRIMARY_ID>
      <SECONDARY_ID>SRR292241</SECONDARY_ID>
    </IDENTIFIERS>
```

Elected Record

This example shows how one record, SRR351940, has replaced 9 others (elected as successor), as in the use case where one run is selected for cSRA loading and the remaining runs are suppressed.

```
<RUN>
    <IDENTIFIERS>
      <PRIMARY_ID>SRR351940</PRIMARY_ID>
      <SECONDARY_ID>SRR351941</SECONDARY_ID>
      <SECONDARY_ID>SRR351942</SECONDARY_ID>
      <SECONDARY_ID>SRR351943</SECONDARY_ID>
      <SECONDARY_ID>SRR351944</SECONDARY_ID>
      <SECONDARY_ID>SRR351945</SECONDARY_ID>
      <SECONDARY_ID>SRR351946</SECONDARY_ID>
      <SECONDARY_ID>SRR351947</SECONDARY_ID>
      <SECONDARY_ID>SRR351948</SECONDARY_ID>
      <SECONDARY_ID>SRR351949</SECONDARY_ID>
   </IDENTIFIERS>
```

Successor Record

This example shows how one record, SRR351940, has replaced another kind of record, analysis object SRZ019522.

```
<RUN>
    <IDENTIFIERS>
      <PRIMARY_ID>SRR351940</PRIMARY_ID>
      <SECONDARY_ID>SRZ019522</SECONDARY_ID>
   </IDENTIFIERS>
```

Submitter alternate identifiers

Submitted records can retain their alternate identifiers and these can be treated as identifiers rather than attributes of the record. The label attribute calls out the display field.

```
<RUN center_name="BI" alias="70291ABXX110301.7.tagged_393.bam" run_center="BI" run_date="2011-03-
01T05:00:00Z" instrument_name="SL-HBZ" accession="SRR404010">
  <IDENTIFIERS>
    <PRIMARY_ID>SRR404010</PRIMARY_ID>
    <SUBMITTER_ID namespace="BI">70291ABXX110301.7.tagged_393.bam</SUBMITTER_ID>
    <SUBMITTER_ID namespace="BI" label="read group platform unit"
>70291ABXX110301.7.CCAGTTAG</SUBMITTER_ID>
  </IDENTIFIERS>
…
```

Submitter replaced identifiers

Submitters can replace an identifier with a new one without disturbing the linkage to existing SRA accessions. However, the primary identifier must be supplied and the defunct identifier must be removed by an update submission.

existing…

```
<RUN alias="454_O.mykiss_GD3412001" accession="SRR090454" center_name="INRA">
<IDENTIFIERS>
    <PRIMARY_ID>SRR090454</PRIMARY_ID>
    <SUBMITTER_ID namespace="INRA">454_O.mykiss_GD3412001</SUBMITTER_ID>
  </IDENTIFIERS>
…
```

updated…

```
<RUN alias="454_O.mykiss_GD3412001" accession="SRR090454" center_name="INRA">
<IDENTIFIERS>
    <PRIMARY_ID>SRR090454</PRIMARY_ID>
    <SUBMITTER_ID namespace="INRA">454_O.mykiss_GB5RBPX02 </SUBMITTER_ID>
  </IDENTIFIERS>
…
```

Commonly used external identifiers

In lieu of a local identifier, a submitter can use a supported external identifier. A good example is a cell line DNA isolate sample from one of the Coriell NA12878:

```
<SAMPLE>
<IDENTIFIERS>
    <PRIMARY_ID >SRR090454</PRIMARY_ID>
    <EXTERNAL_ID namespace="Coriell" label="Catalog ID">NA12878 </EXTERNAL_ID>
    <EXTERNAL_ID namespace="Coriell" label="Catalog ID">GM12878 </EXTERNAL_ID>
  </IDENTIFIERS>
```

…

External identifiers must use a namespace attribute that is registered with the SRA archive.

Universally unique identifiers

A downstream user of SRA xml data could annotate it with a universally unique identifier. This requires no namespace because it is universally unique (according to the generation method). The INSDC SRAs do not use UUIDs and these are ignored on submission.

```
<RUN alias="68b329da9893e34099c7d8ad5cb9c940" accession="SRR090454" center_name="">
<IDENTIFIERS>
    <PRIMARY_ID>SRR090454</PRIMARY_ID>
    <UUID> 68b329da9893e34099c7d8ad5cb9c940 </UUID>
  </IDENTIFIERS>
…
```