<!-- ============================================ -->

<!-- This section mapped from ASN.1 module INSD-INSDSeq -->

<!-- ============================================ -->

<!-- Definition of INSDSet -->

<!--

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

ASN.1 and XML for the components of a GenBank/EMBL/DDBJ sequence record

The International Nucleotide Sequence Database (INSD) collaboration

Version 1.5, 15 January 2009

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

INSDSeq provides the elements of a sequence as presented in the

GenBank/EMBL/DDBJ-style flatfile formats, with a small amount of

additional structure.

Although this single perspective of the three flatfile formats

provides a useful simplification, it hides to some extent the

details of the actual data underlying those formats. Nevertheless,

the XML version of INSD-Seq is being provided with

the hopes that it will prove useful to those who bulk-process

sequence data at the flatfile-format level of detail. Further

documentation regarding the content and conventions of those formats

can be found at:

URLs for the DDBJ, EMBL, and GenBank Feature Table Document:

http://www.ddbj.nig.ac.jp/FT/full\_index.html

http://www.ebi.ac.uk/embl/Documentation/FT\_definitions/feature\_table.html

http://www.ncbi.nlm.nih.gov/projects/collab/FT/index.html

URLs for DDBJ, EMBL, and GenBank Release Notes :

ftp://ftp.ddbj.nig.ac.jp/database/ddbj/ddbjrel.txt

http://www.ebi.ac.uk/embl/Documentation/Release\_notes/current/relnotes.html

ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

Because INSDSeq is a compromise, a number of pragmatic decisions have

been made:

In pursuit of simplicity and familiarity a number of fields do not

have full substructure defined here where there is already a

standard flatfile format string. For example:

Dates: DD-MON-YYYY (eg 10-JUN-2003)

Author: LastName, Initials (eg Smith, J.N.)

or Lastname Initials (eg Smith J.N.)

Journal: JournalName Volume (issue), page-range (year)

or JournalName Volume(issue):page-range(year)

eg Appl. Environ. Microbiol. 61 (4), 1646-1648 (1995)

Appl. Environ. Microbiol. 61(4):1646-1648(1995).

FeatureLocations are representated as in the flatfile feature table,

but FeatureIntervals may also be provided as a convenience

FeatureQualifiers are represented as in the flatfile feature table.

Primary has a string that represents a table to construct

a third party (TPA) sequence.

other-seqids can have strings with the "vertical bar format" sequence

identifiers used in BLAST for example, when they are non-INSD types.

Currently in flatfile format you only see Accession numbers, but there

are others, like patents, submitter clone names, etc which will

appear here

There are also a number of elements that could have been more exactly

specified, but in the interest of simplicity have been simply left as

optional. For example:

All publicly accessible sequence records in INSDSeq format will

include accession and accession.version. However, these elements are

optional in optional in INSDSeq so that this format can also be used

for non-public sequence data, prior to the assignment of accessions and

version numbers. In such cases, records will have only "other-seqids".

sequences will normally all have "sequence" filled in. But contig records

will have a "join" statement in the "contig" slot, and no "sequence".

We also may consider a retrieval option with no sequence of any kind

and no feature table to quickly check minimal values.

Four (optional) elements are specific to records represented via the EMBL

sequence database: INSDSeq\_update-release, INSDSeq\_create-release,

INSDSeq\_entry-version, and INSDSeq\_database-reference.

One (optional) element is specific to records originating at the GenBank

and DDBJ sequence databases: INSDSeq\_segment.

\*\*\*\*\*\*\*\*

-->

<!-- Entities used to give specificity to #PCDATA -->

<!ENTITY % INTEGER '#PCDATA'>

<!ENTITY % ENUM 'EMPTY'>

<!ENTITY % BOOLEAN 'EMPTY'>

<!ENTITY % NULL 'EMPTY'>

<!ENTITY % REAL '#PCDATA'>

<!ENTITY % OCTETS '#PCDATA'>

<!ELEMENT INSDSet ( INSDSeq\* )>

<!-- Definition of INSDSeq -->

<!ELEMENT INSDSeq (

INSDSeq\_locus? ,

INSDSeq\_length ,

INSDSeq\_strandedness? ,

INSDSeq\_moltype ,

INSDSeq\_topology? ,

INSDSeq\_division? ,

INSDSeq\_update-date? ,

INSDSeq\_create-date? ,

INSDSeq\_update-release? ,

INSDSeq\_create-release? ,

INSDSeq\_definition? ,

INSDSeq\_primary-accession? ,

INSDSeq\_entry-version? ,

INSDSeq\_accession-version? ,

INSDSeq\_other-seqids? ,

INSDSeq\_secondary-accessions? ,

INSDSeq\_project? ,

INSDSeq\_keywords? ,

INSDSeq\_segment? ,

INSDSeq\_source? ,

INSDSeq\_organism? ,

INSDSeq\_taxonomy? ,

INSDSeq\_references? ,

INSDSeq\_comment? ,

INSDSeq\_comment-set? ,

INSDSeq\_struc-comments? ,

INSDSeq\_primary? ,

INSDSeq\_source-db? ,

INSDSeq\_database-reference? ,

INSDSeq\_feature-table? ,

INSDSeq\_feature-set? ,

INSDSeq\_sequence? ,

INSDSeq\_contig? ,

INSDSeq\_alt-seq? ,

INSDSeq\_xrefs? )>

<!ELEMENT INSDSeq\_locus ( #PCDATA )>

<!ELEMENT INSDSeq\_length ( %INTEGER; )>

<!ELEMENT INSDSeq\_strandedness ( #PCDATA )>

<!ELEMENT INSDSeq\_moltype ( #PCDATA )>

<!ELEMENT INSDSeq\_topology ( #PCDATA )>

<!ELEMENT INSDSeq\_division ( #PCDATA )>

<!ELEMENT INSDSeq\_update-date ( #PCDATA )>

<!ELEMENT INSDSeq\_create-date ( #PCDATA )>

<!ELEMENT INSDSeq\_update-release ( #PCDATA )>

<!ELEMENT INSDSeq\_create-release ( #PCDATA )>

<!ELEMENT INSDSeq\_definition ( #PCDATA )>

<!ELEMENT INSDSeq\_primary-accession ( #PCDATA )>

<!ELEMENT INSDSeq\_entry-version ( #PCDATA )>

<!ELEMENT INSDSeq\_accession-version ( #PCDATA )>

<!ELEMENT INSDSeq\_other-seqids ( INSDSeqid\* )>

<!ELEMENT INSDSeq\_secondary-accessions ( INSDSecondary-accn\* )>

<!--

INSDSeq\_project has been deprecated in favor of INSDSeq\_xrefs .

This element may be be removed from a future version of this DTD.

-->

<!ELEMENT INSDSeq\_project ( #PCDATA )>

<!ELEMENT INSDSeq\_keywords ( INSDKeyword\* )>

<!ELEMENT INSDSeq\_segment ( #PCDATA )>

<!ELEMENT INSDSeq\_source ( #PCDATA )>

<!ELEMENT INSDSeq\_organism ( #PCDATA )>

<!ELEMENT INSDSeq\_taxonomy ( #PCDATA )>

<!ELEMENT INSDSeq\_references ( INSDReference\* )>

<!ELEMENT INSDSeq\_comment ( #PCDATA )>

<!ELEMENT INSDSeq\_comment-set ( INSDComment\* )>

<!ELEMENT INSDSeq\_struc-comments ( INSDStrucComment\* )>

<!ELEMENT INSDSeq\_primary ( #PCDATA )>

<!ELEMENT INSDSeq\_source-db ( #PCDATA )>

<!ELEMENT INSDSeq\_database-reference ( #PCDATA )>

<!ELEMENT INSDSeq\_feature-table ( INSDFeature\* )>

<!ELEMENT INSDSeq\_feature-set ( INSDFeatureSet\* )>

<!--

Optional for contig, wgs, etc.

-->

<!ELEMENT INSDSeq\_sequence ( #PCDATA )>

<!ELEMENT INSDSeq\_contig ( #PCDATA )>

<!ELEMENT INSDSeq\_alt-seq ( INSDAltSeqData\* )>

<!--

INSDSeq\_xrefs provides cross-references from a sequence record

to other database resources. These cross-references are at the

level of the entire record, rather than at the level of a specific

feature. These cross-references can include: BioProject, BioSample,

Sequence Read Archive, etc.

-->

<!ELEMENT INSDSeq\_xrefs ( INSDXref\* )>

<!-- Definition of INSDSeqid -->

<!ELEMENT INSDSeqid ( #PCDATA )>

<!-- Definition of INSDSecondary-accn -->

<!ELEMENT INSDSecondary-accn ( #PCDATA )>

<!-- Definition of INSDKeyword -->

<!--

INSDReference\_position contains a string value indicating the

basepair span(s) to which a reference applies. The allowable

formats are:

X..Y : Where X and Y are integers separated by two periods,

X >= 1 , Y <= sequence length, and X <= Y

Multiple basepair spans can exist, separated by a

semi-colon and a space. For example : 10..20; 100..500

sites : The string literal 'sites', indicating that a reference

provides sequence annotation information, but the specific

basepair spans are either not captured, or were too numerous

to record.

The 'sites' literal string is singly occuring, and

cannot be used in conjunction with any X..Y basepair spans.

'sites' is a convention utilized by GenBank, and might

not be presented in XML provided by EMBL and DDBJ.

References that lack an INSDReference\_position element are not

attributed to any particular region of the sequence.

-->

<!ELEMENT INSDKeyword ( #PCDATA )>

<!-- Definition of INSDReference -->

<!ELEMENT INSDReference (

INSDReference\_reference ,

INSDReference\_position? ,

INSDReference\_authors? ,

INSDReference\_consortium? ,

INSDReference\_title? ,

INSDReference\_journal ,

INSDReference\_xref? ,

INSDReference\_pubmed? ,

INSDReference\_remark? )>

<!ELEMENT INSDReference\_reference ( #PCDATA )>

<!ELEMENT INSDReference\_position ( #PCDATA )>

<!ELEMENT INSDReference\_authors ( INSDAuthor\* )>

<!ELEMENT INSDReference\_consortium ( #PCDATA )>

<!ELEMENT INSDReference\_title ( #PCDATA )>

<!ELEMENT INSDReference\_journal ( #PCDATA )>

<!ELEMENT INSDReference\_xref ( INSDXref\* )>

<!ELEMENT INSDReference\_pubmed ( %INTEGER; )>

<!ELEMENT INSDReference\_remark ( #PCDATA )>

<!-- Definition of INSDAuthor -->

<!ELEMENT INSDAuthor ( #PCDATA )>

<!-- Definition of INSDXref -->

<!--

INSDXref provides a method for referring to records in

other databases. INSDXref\_dbname is a string value that

provides the name of the database, and INSDXref\_dbname

is a string value that provides the record's identifier

in that database.

-->

<!ELEMENT INSDXref (

INSDXref\_dbname ,

INSDXref\_id )>

<!ELEMENT INSDXref\_dbname ( #PCDATA )>

<!ELEMENT INSDXref\_id ( #PCDATA )>

<!-- Definition of INSDComment -->

<!ELEMENT INSDComment (

INSDComment\_type? ,

INSDComment\_paragraphs )>

<!ELEMENT INSDComment\_type ( #PCDATA )>

<!ELEMENT INSDComment\_paragraphs ( INSDCommentParagraph\* )>

<!-- Definition of INSDCommentParagraph -->

<!ELEMENT INSDCommentParagraph ( #PCDATA )>

<!-- Definition of INSDStrucComment -->

<!ELEMENT INSDStrucComment (

INSDStrucComment\_name? ,

INSDStrucComment\_items )>

<!ELEMENT INSDStrucComment\_name ( #PCDATA )>

<!ELEMENT INSDStrucComment\_items ( INSDStrucCommentItem\* )>

<!-- Definition of INSDStrucCommentItem -->

<!ELEMENT INSDStrucCommentItem (

INSDStrucCommentItem\_tag? ,

INSDStrucCommentItem\_value? ,

INSDStrucCommentItem\_url? )>

<!ELEMENT INSDStrucCommentItem\_tag ( #PCDATA )>

<!ELEMENT INSDStrucCommentItem\_value ( #PCDATA )>

<!ELEMENT INSDStrucCommentItem\_url ( #PCDATA )>

<!-- Definition of INSDFeatureSet -->

<!--

INSDFeature\_operator contains a string value describing

the relationship among a set of INSDInterval within

INSDFeature\_intervals. The allowable formats are:

join : The string literal 'join' indicates that the

INSDInterval intervals are biologically joined

together into a contiguous molecule.

order : The string literal 'order' indicates that the

INSDInterval intervals are in the presented

order, but they are not necessarily contiguous.

Either 'join' or 'order' is required if INSDFeature\_intervals

is comprised of more than one INSDInterval .

-->

<!ELEMENT INSDFeatureSet (

INSDFeatureSet\_annot-source? ,

INSDFeatureSet\_features )>

<!ELEMENT INSDFeatureSet\_annot-source ( #PCDATA )>

<!ELEMENT INSDFeatureSet\_features ( INSDFeature\* )>

<!-- Definition of INSDFeature -->

<!ELEMENT INSDFeature (

INSDFeature\_key ,

INSDFeature\_location ,

INSDFeature\_intervals? ,

INSDFeature\_operator? ,

INSDFeature\_partial5? ,

INSDFeature\_partial3? ,

INSDFeature\_quals? ,

INSDFeature\_xrefs? )>

<!ELEMENT INSDFeature\_key ( #PCDATA )>

<!ELEMENT INSDFeature\_location ( #PCDATA )>

<!ELEMENT INSDFeature\_intervals ( INSDInterval\* )>

<!ELEMENT INSDFeature\_operator ( #PCDATA )>

<!ELEMENT INSDFeature\_partial5 %BOOLEAN; >

<!ATTLIST INSDFeature\_partial5 value ( true | false ) #REQUIRED >

<!ELEMENT INSDFeature\_partial3 %BOOLEAN; >

<!ATTLIST INSDFeature\_partial3 value ( true | false ) #REQUIRED >

<!ELEMENT INSDFeature\_quals ( INSDQualifier\* )>

<!ELEMENT INSDFeature\_xrefs ( INSDXref\* )>

<!-- Definition of INSDInterval -->

<!--

-->

<!--

INSDInterval\_iscomp is a boolean indicating whether

an INSDInterval\_from / INSDInterval\_to location

represents a location on the complement strand.

When INSDInterval\_iscomp is TRUE, it essentially

confirms that a 'from' value which is greater than

a 'to' value is intentional, because the location

is on the opposite strand of the presented sequence.

INSDInterval\_interbp is a boolean indicating whether

a feature (such as a restriction site) is located

between two adjacent basepairs. When INSDInterval\_interbp

is TRUE, the 'from' and 'to' values will differ by

exactly one base for linear molecules. For circular

molecules, if the inter-basepair position falls between

the last and the first base, then 'from' will be the

final base (equal to the length of the sequence), and

'to' will have a value of 1 .

-->

<!ELEMENT INSDInterval (

INSDInterval\_from? ,

INSDInterval\_to? ,

INSDInterval\_point? ,

INSDInterval\_iscomp? ,

INSDInterval\_interbp? ,

INSDInterval\_accession )>

<!ELEMENT INSDInterval\_from ( %INTEGER; )>

<!ELEMENT INSDInterval\_to ( %INTEGER; )>

<!ELEMENT INSDInterval\_point ( %INTEGER; )>

<!ELEMENT INSDInterval\_iscomp %BOOLEAN; >

<!ATTLIST INSDInterval\_iscomp value ( true | false ) #REQUIRED >

<!ELEMENT INSDInterval\_interbp %BOOLEAN; >

<!ATTLIST INSDInterval\_interbp value ( true | false ) #REQUIRED >

<!ELEMENT INSDInterval\_accession ( #PCDATA )>

<!-- Definition of INSDQualifier -->

<!ELEMENT INSDQualifier (

INSDQualifier\_name ,

INSDQualifier\_value? )>

<!ELEMENT INSDQualifier\_name ( #PCDATA )>

<!ELEMENT INSDQualifier\_value ( #PCDATA )>

<!--

INSDAltSeqData provides for sequence representations other than

literal basepair abbreviations (INSDSeq\_sequence), such as the

CONTIG/CO linetype of the GenBank and EMBL flatfile formats.

It also accomodates the specification of accession-number ranges,

which are presented on a WGS master record (for the contigs and

and scaffolds of a WGS project).

-->

<!-- Definition of INSDAltSeqData -->

<!ELEMENT INSDAltSeqData (

INSDAltSeqData\_name ,

INSDAltSeqData\_items? )>

<!--

e.g., contig, wgs, scaffold, cage, genome

-->

<!ELEMENT INSDAltSeqData\_name ( #PCDATA )>

<!ELEMENT INSDAltSeqData\_items ( INSDAltSeqItem\* )>

<!-- Definition of INSDAltSeqItem -->

<!ELEMENT INSDAltSeqItem (

INSDAltSeqItem\_interval? ,

INSDAltSeqItem\_isgap? ,

INSDAltSeqItem\_gap-length? ,

INSDAltSeqItem\_gap-type? ,

INSDAltSeqItem\_gap-linkage? ,

INSDAltSeqItem\_gap-comment? ,

INSDAltSeqItem\_first-accn? ,

INSDAltSeqItem\_last-accn? ,

INSDAltSeqItem\_value? )>

<!ELEMENT INSDAltSeqItem\_interval ( INSDInterval )>

<!ELEMENT INSDAltSeqItem\_isgap %BOOLEAN; >

<!ATTLIST INSDAltSeqItem\_isgap value ( true | false ) #REQUIRED >

<!ELEMENT INSDAltSeqItem\_gap-length ( %INTEGER; )>

<!ELEMENT INSDAltSeqItem\_gap-type ( #PCDATA )>

<!ELEMENT INSDAltSeqItem\_gap-linkage ( #PCDATA )>

<!ELEMENT INSDAltSeqItem\_gap-comment ( #PCDATA )>

<!ELEMENT INSDAltSeqItem\_first-accn ( #PCDATA )>

<!ELEMENT INSDAltSeqItem\_last-accn ( #PCDATA )>

<!ELEMENT INSDAltSeqItem\_value ( #PCDATA )>