XML for BLAST

The NCBI is now making a new version of XML available for BLAST. This new version includes better support for searches against many queries at once, taxonomic information, non-redundant databases, and XML Schema. The rest of this document lists many of the changes in the new XML version.

Currently, BLAST results using the new XML can be obtained from the NCBI BLAST website at blast.ncbi.nlm.nih.gov with the following steps. First, run a BLAST search at blast.ncbi.nlm.nih.gov. Second, view the standard report for your search. Third, open the "Download" link at the top of the page. Finally, select XML2 to download a zip file with your results in the new XML. Instead of XML2, you may select JSON to see your results in JSON format using the new XML specification. Stand-alone BLAST executables that produce the new XML should be available soon. You may use provide feedback at http://www.ncbi.nlm.nih.gov/sites/ehelp/feedback?Ncbi_App=blastXML&Data=JiraApp:HD

One query per xml file:-

Instead of one XML document for all queries in a search, the result for each query is written to a separate XML file. Each result file is labeled with the user supplied filename (or RID if from the web) followed by an underscore and a number which corresponds to the position of the query in the input query file.

Example: test_1.xml (Result for Query 1)

Xinclude file:-

The new BLAST XML (BLAST XML2) will produce an XInclude file which can be used to generate a single XML document that contains results from all the queries in a search. The XInclude file is labeled with just the user supplied filename (or RID).

Example: test.xml (XInclude File)

Prefix removed for child elements:

Prefixes for child elements have been removed.

Example:-

<program>blastn<program>   --- BLAST XML2

<BlastOutput_program>blastn<BlastOutput_program>   -- old XML

XML associated with schema:-

The schema location (below) is included in the XML file. It can be used for validation http://www.ncbi.nlm.nih.gov/data_specs/schema_alt/NCBI_BlastOutput2.xsd

Structural Changes – Data Reorganization:

The top level BLAST XML2 element <BlastOutput2> contains two child elements, <Err> and <Report>. <Err> is for storing an error message and an error code if an error occurs during the search. <Report> captures the content of a BLAST report for a single query. It contains all the elements that go under <BlastOutput> in the old XML format. Figure 1 and Figure 2 below show the top level overview of the old and new BLAST XML schemas.

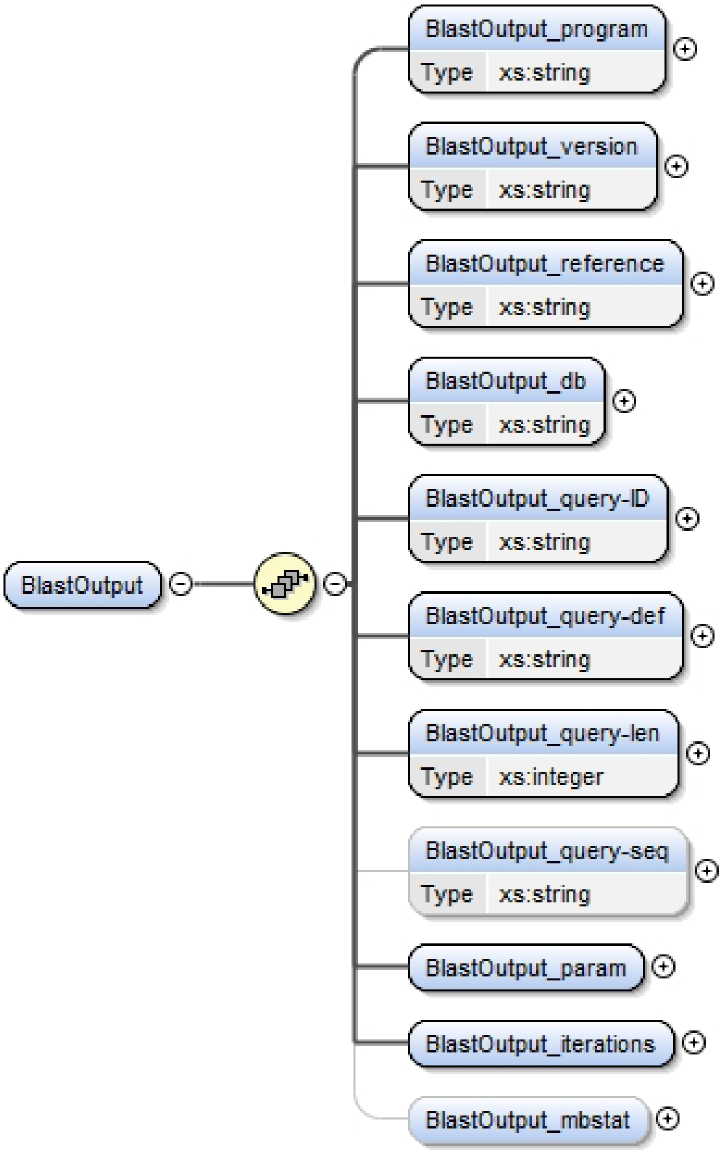

Figure 1: BLAST XML (old) Top Level Schema Diagram

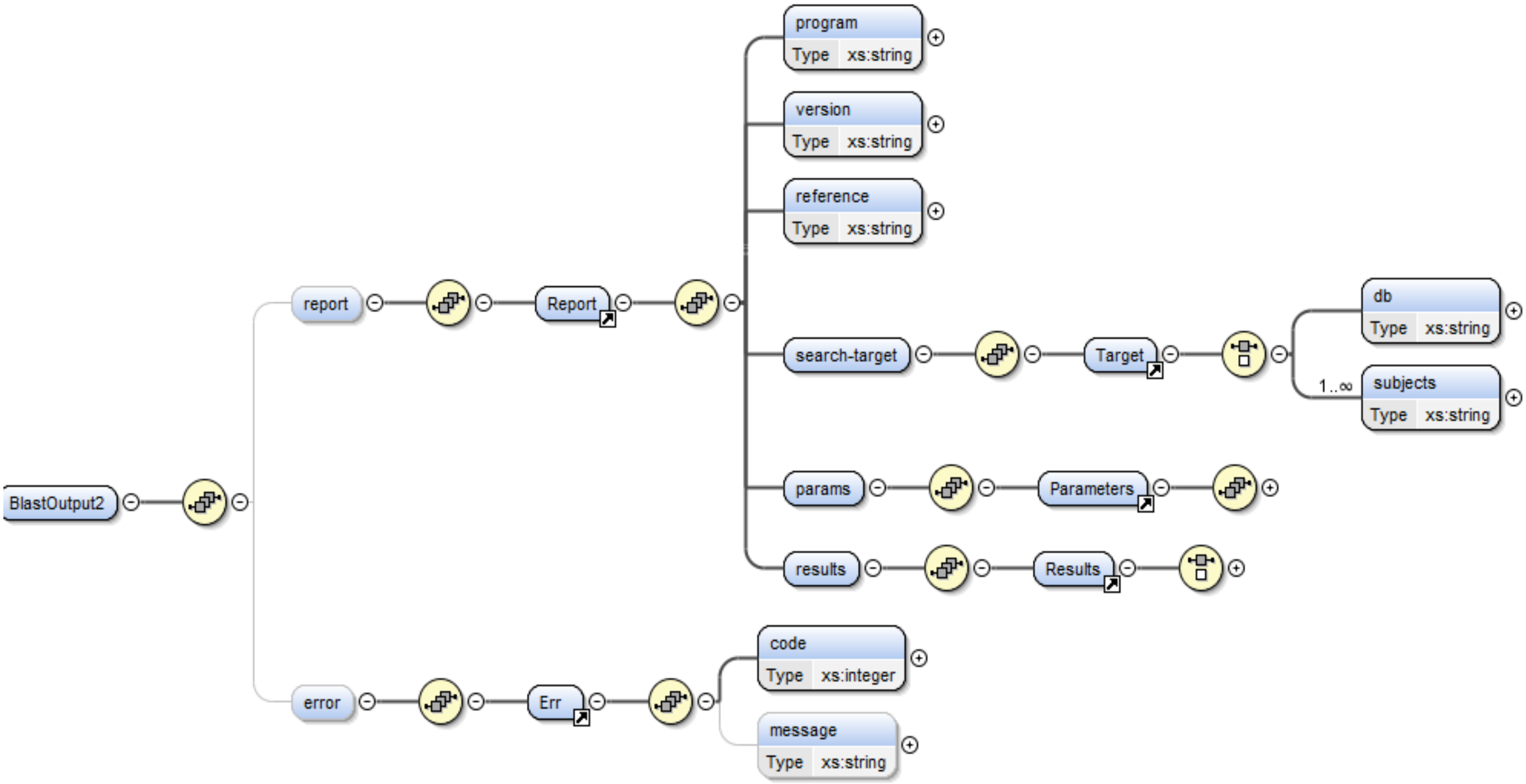

Figure 2: BLAST XML2 Top Level Schema Diagram

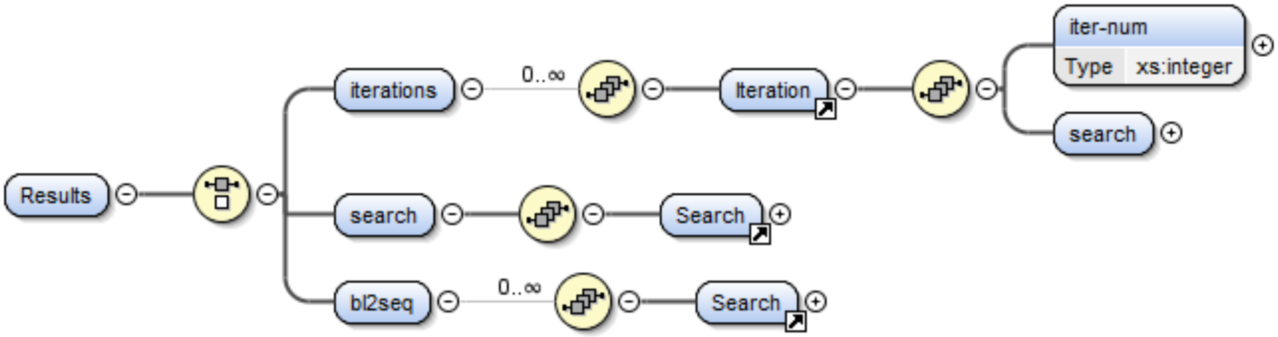

Figure 3: BLAST XML2 <Results>

Query information (<query-ID>, <query-def>, <query-len> and <query-seq>) have been removed from under the top level element in Blast XML2.

A new element <subjects> has been added under a new element called <search-target> to support Blast2seq mode. The element <db>, which is a child of the top level element <BlastOuput> in old XML has been placed under <search-target> in the BLAST XML2 format.

As shown in Figure 2 and Figure 3, the element <Results> in the BLAST XML2 format has replaced <BlastOutput_Iteration> in the original format. <Results> can contain any one of the three elements <iteration>, <search> and <bl2seq>, which are mutually exclusive.

<iteration> captures the results for iterative blast searches such as psi-blast and delta-blast, it is a container for a series of <Iteration> and each of this element in the series corresponds to the result of one single iteration.

<search> is used for non-iterative database searches.

<bl2seq> is used for bl2seq searches, each <Search> under <bl2seq> represents the result of one query-subject pair, the order of the results is presented according to the order of subjects in the subject file (note that the number of elements in <bl2seq> is always the same as the number of subjects).

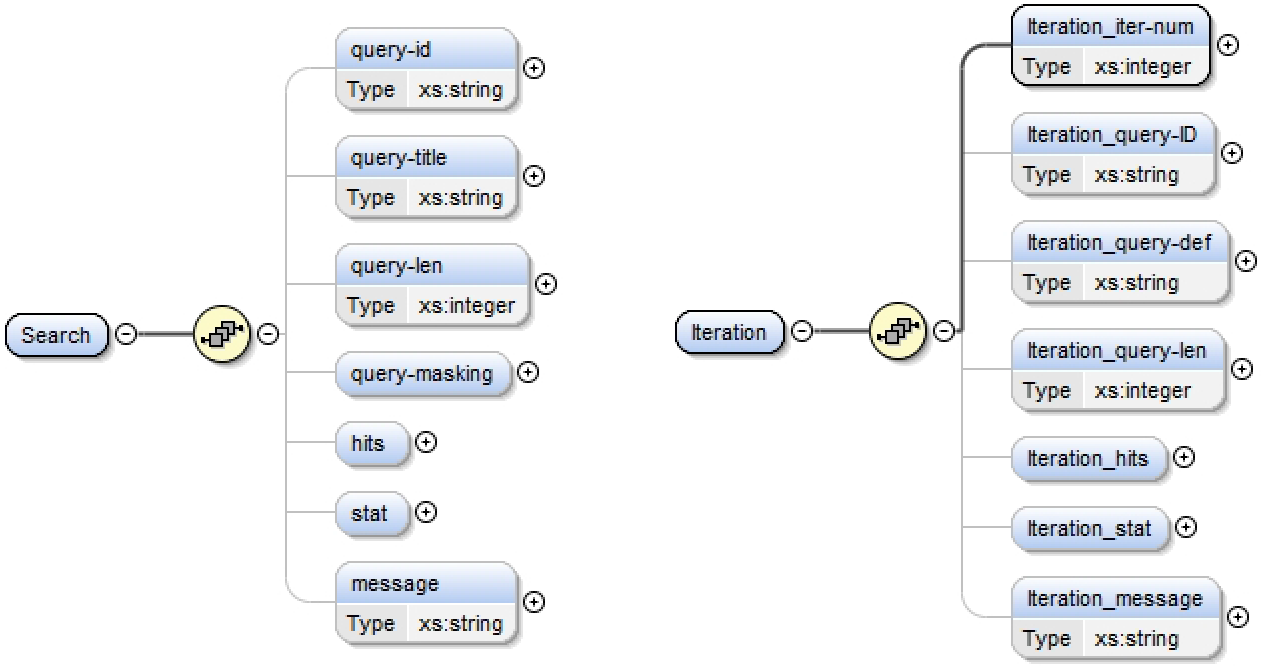

Figure 4: Blast XML2 <Search> (on the left) and Old XML <Iteration> (on the right)

As shown in Figure 4, most of the elements originally fall under <Iteration> have been reorganized under <Search>, which is the building block for <search>, <iteration> and <bl2seq> as described above.

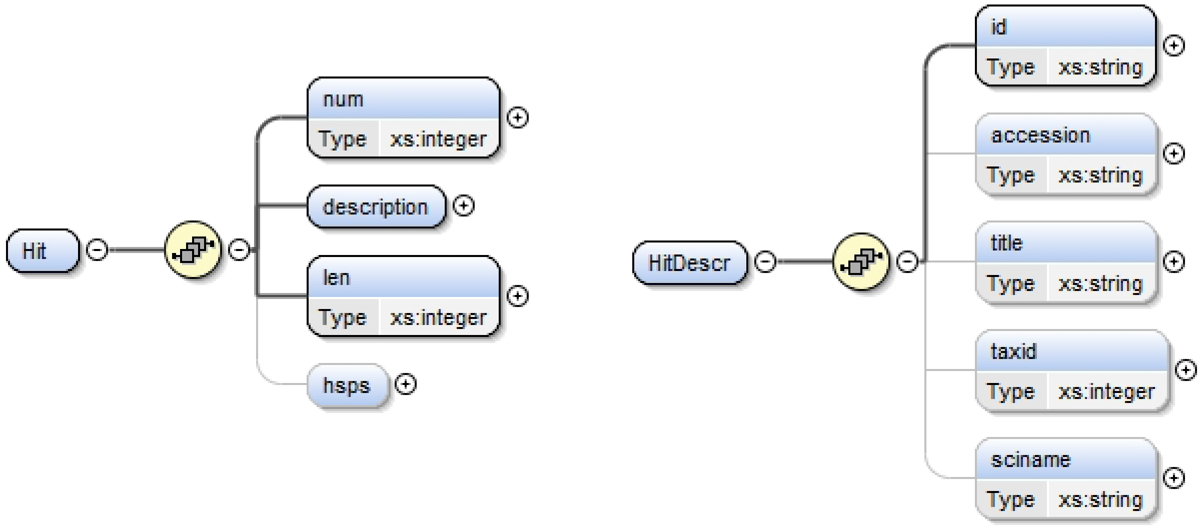

Figure 5: Blast XML2 <Hit> and <HitDescr>

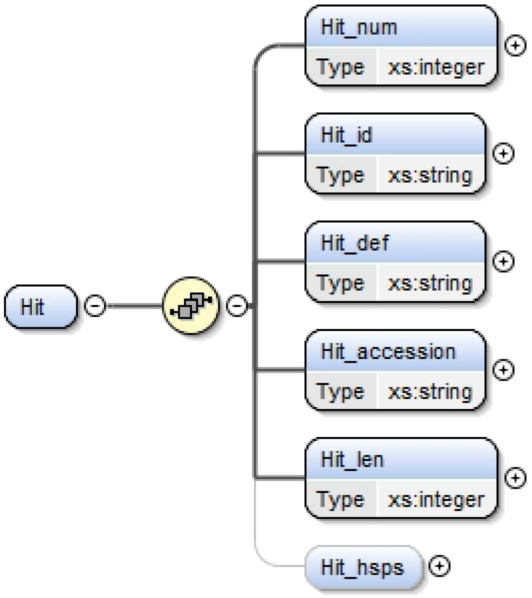

Figure 6: Old Blast XML <Hit>

One entry (sequence) in the BLAST database may correspond to multiple sequence identifiers. In the old XML format, only the first sequence identifier has its own <Hit_id> and <Hit_accession> elements, but the rest of the entries are all concatenated into the <Hit_def> element, along with the title for the first identifier. In BLAST XML2, each sequence identifier will have its own <HitDescr> element as shown in

Figure 5. Additionally, each <HitDescr> contains the title, NCBI taxid and the scientific name (genus species) for that sequence identifier.

New Non-Structural Elements:-

Name	Child Of Element	Type	Description
cbs	Parameters	integer	Composition-based statistics
query-gencode	Parameters	integer	Gencode used for translating query
db-gencode	Parameters	integer	Gencode used for translating subject
query-masking	Search	Range	A list of query masking locations
Range	N/A	Pair of integer	Range defined by from and to integer pair
query-strand	Hsp	string	Query strand (Plus/Minus)
hit-strand	Hsp	string	Subject Strand (Plus/Minus)
taxid	HitDescr	integer	NCBI taxonomy ID
sciname	HitDescr	string	Binomial scientific name

Name changes:

Old XML		XML2	
Parent	Element	Parent	Element
Iteration	query-def	Search	query-title
Iteration	query-ID	Search	query-id
Hit	def	HitDescr	title

Type Changes:-

Parent	Element	Old Type	XML2 Type
Statistics	eff-space	double	long
Statisitics	db-num	integer	long