# SPL - Layered Hash

Layered hash - document identifier string - will be used to identify a document and locate similar documents.

We need to understand how users of SRS would want to locate documents with specific substance(s). What search terms are relevant for locating substances? We will use this knowledge to define a structure of a SPL document identifier string.

### Document data store search API.

* has\_polypeptide\_chain( chain\_id )
  + chain\_id - (hash) code representing a protein sequence.
  + return collection of documents that contain specified chain.
* has\_chemical\_structure( content\_type, content\_value )
  + Parameters:
    - content\_type: molfile, InChi, InChiKey.
    - content\_value : a string representing value of a chemical structure. Molfile content will be represented by its hash value.
  + Return collection of documents that contain specified chemical structure.
* has\_structural\_modifications( chain\_id, mod\_type, positions)
  + Parameters:
    - chain\_id : polypeptide chain id.
    - mod\_type : [ substitution, attachment ].
    - position: list of positions on chain sequence.
  + Return collection of documents that contain specified chain that have structural modifications of specified type for specified positions. If positions list is empty, then do not filter result set by positions.
* has\_substitutions( chain\_id, chemical\_structure, positions)
  + Parameters:
    - chain\_id : polypeptide chain id.
    - chemical\_structure : chemical structure descriptor such as ( content\_type=[molfile, InChi, InChiKey], content).
    - positions: list of positions on polypeptide chain. If positions list is empty then do not filter result set by positions.
* has\_bridge( chain\_id, chemical\_struct )
  + Parameters:
    - chain\_id : polypeptide chain id.
    - chemical\_struct : chemical structure description ( content\_type, content ) of a link molecule.
  + Return documents that contain specified polypeptide chain such that there is(are) intra - chain bridge(s) with specified link molecule.
* has\_bridge( chain\_id, chain\_id, chemical\_struct )
  + Parameters:
    - chain\_id : polypeptide chain id that serves as a bridge end.
    - chain\_id : polypeptide chain id that serves as a bridge end.
    - chemical\_structure: chemical structure description ( content\_type, content ) of a link molecule.
  + Return documents that contain specified chains such that the chains are linked by specified molecule.

### Document data store structure.

A data store will have following collections to provide ability to search for SPL documents:

* Documents collection will store (key, document id) pairs where key values correspond to identifier strings derived from a document. This collection would be used to locate similar / duplicate documents.
* Documents collection
  + key
  + SPL document id

Examples

( “/chain=A”, docid ), # Chain with id “A” => document id

( “/mod=A:sub” , docid ), # Substitution on chain with id “A” => document id

(“/mod=A:sub:33”, docid ), # Substitution on chain with id “A” at position 33 => document

( “/polymer=P”, docid ), # substance with encoded chemical struct => document

( “/polymer=P:CP”, docid ) # substance with encoded chemical struct + connection points=> document

( “/sub=A:P”, docid ) # AA on chain with id “A” is substituted by a molecule with id “P” => document

( “/sub=A:33:P”, docid ) # AA on chain “A” at position “33” is substituted by a molecule “P” => document

( “/link=A:P“, docid ) # intra-chain bridge between regions of a chain with id “A”, link molecule id “P” => document

( “/link=A:A:P”, docid ) # inter-chain bridge between identical chains with id “A”, link molecule id “P” => document.

### Layers of SPL document identifier string.

A SPL document identifier string will consist of parts / layers that capture document’s content at different level of detail.

Layers.

* Chains layer.
  + Chains layer represents collection of polypeptide chain identifiers. A chain identifier is computed from a polypeptide amino acid sequence. The layer will contain duplicate identifiers if substance description contains two or more identical chains. E.g. “/chains=A;A;B;C;” string defines identifiers of four chains of a substance, where two chains have identical AA sequence.
* Polymer (need a better name) layer.
  + Polymer (need a better name) layer represents chemical structure(s) that modify main substance. Identifier string for a polymer is a compound identifier: one part is computed from its chemical structure content - InChiKey, or InChi, or Molfile and the other part encodes collection of its connection points. E.g. “/polymer=P:abc;M:xyz”, where ‘P' and ‘M’ represent encoded chemical structure value and ‘abc’ and 'xyz’ represent encoded connection points collections.
* Modified chains layer.
  + Modified chains layer represents collection of mappings from chain identifier to modified chain identifiers. Modified chain identifier string captures modifications and substitutions for the chain. E.g. “/modified\_chains=A:X:Y;B:Z;C:W” identifier string defines that two instances of chain ‘A' have ‘X’ and 'Y’ identifier strings assigned as a result of modifications.
* Modifications layer.
  + Modifications layer identifier captures chains' structural modifications that can be either substitutions or attachments. E.g. /modifications=A[1]:sub:23;A[1]:sub:33;A[1]:sub:144;A[2]:sub:23;A[2]:sub:133;A[2]:sub:144”
* Substitutions layer.
  + Substitutions layer describes substitutions in polypeptide chains, where a substitution captures chain id, position in a chain’s sequence, polymer identifier and its connection point index. E.g. “/substitutions=A[1]:33:P:1”.
* Links layer.
  + Links layer describes intra- and inter-chain bridges. E.g. “/links=A[1]:23:A[1]:144;A[1]:33:A[2]:133” defines two bridges: intra-chain bridge for chain A[1] from connecting positions 1 and 144, and inter-chain bridge, connecting chains A[1] and A[2], positions 33 and 133 respectively.A concrete bridge’s link molecule and its connection points is defined by “/substitutions” layer.

### Formal definition of identifier string structure.

C:\e7344b195eb55adf1e9f2d58c85fc0a8Click here to expand...

1.1 Polypeptide chains layer.

Polypeptide chains layer consist of a prefix string "/chains="

and ';' separated collection of chain identifiers:

<chain-layer> ::= "/chains=" <chain-id> [ ";" <chain-id> ]\*

<chain-id> ::= Hash( Canonical( <aa-sequence> ) )

<aa-sequence> ::= Amino Acid Sequence

Hash :: Sha256 or MD5 cryptographic hash function.

Canonical ::= (A: string): { return A if A < reverse(A) else return reverse(A); }

1.2. Modified chains layer.

<modified-chains-layer> ::= "/modified\_chains=" <chain-id> ":" <chain-instance> [ "," <chain-instance ]\*

<chain-instance> ::= hash code of a modified chan.

1.3. Modifications layer.

Modification of a polypeptide chain layer consists of a prefix string

"/modifications=" and ";" separated collection of modification identifiers.

<modification-layer> ::= "/modifications=" <modification-id> [ ";" <modification-id> ]\*

<modification-id> ::= <modification-type> ":" <chain-instance-id> ":" <chain-position>

<modification-type> ::= <aa-substitution> | <structural-attachment`>

<chain-instance-id> ::= <chain-id> | <chain-id> "[" integer "]"

<aa-substitution> ::= "sub" # "moiety[code=C118425]/partMoiety/bond[code=C118426]"

<structural-attachment> ::= "attach" # "moiety[code=C118425]/partMoiety/bond[code=C14050]"

<chain-position> ::= <position> # Amino acid position on polypeptide chain

<position> ::= integer | "NA"

Elements representing structural modification of a poypeptide chains can be located in SPL XML document with following XPath expression:

"//moiety[code[@code = 'C118425' ]][descendant::bond[code[@code = 'C118426']]"

"//moiety[code[@code = 'C118425' ]][descendant::bond[code[@code = 'C14050']]"

1.4 Polymer layer.

Polymer layer represents substances that replace an amino acid in a polypeptide chain. Polymer layer consists of a prefix "/polymer=" and ';' separated

collection of polymer identifiers.

<polymer-layer> ::= "/polymer=" <polymer-id> [ ";" <polymer-id> ]\*

<polymer-id> ::= Hash( <chemical-struct> )

<chemical-struct> ::= <inchi-key> | <inchi> | <molfile>

<inchi-key> ::= "/inchikey=" <value>

<inchi> ::= "/inchi=" <value>

<molfile> ::= "/molfile=" <value>

<value> ::= XPath("/subjectOd/characteristic/value/@text")

1.5 Substitutions layer.

Substitution layer represents concrete instances of

structural modifications - amino acid substitutions.

<substition-layer> ::= "/substitution=" <substitution-id> [ ";" <substitution-id> ]\*

<substitution-id> ::= <modification-id> ":" <polymer-id> ":" <connection-point> ":" <index>

<connection-points> ::= Hash ( <N> "-" <C> [ ";" <N> "-" <C> ]\* )

<index> ::= integer # 1-based index within conn points collection

<N> ::= integer # Position of N atoms

<C> ::= integer # Position of C atoms

1.6. Link layer.

Link layer captures end points of a bridge structure.

<link-layer> ::= "/link=" <bridge> [ ";" <bridge> ]\*

<bridge> ::= <chain-instance-id> ":" <position> ":" <chain-instance-id> ":" <position>

<position> ::= <cinteger>

1.7. Quantity layer.

Quantity layer represents amount of substitution.

<quantity-layer> ::= "/quantity=" <quantity-id> [ "," <quantity-id> ]\*

<quantity-id> ::= <substitution-id> ":" <numerator> ":" <denominator>

<numerator> ::= <amount> ":" <unit>

<denomintator> ::= <amount> ":" <unit>

<amount> ::= integer

<unit> ::= "mol"

Q&A.

If we only go by the sequence, because it is identical for both chains it would be not clear if the link links different positions on the same chain or links to a different chain. Does that make sense?

We will use subscript [] operator to specify concrete instance of a chain id, when two or more chains have identical sequence. “Modified chains” layer would define association between chain id and modified instance(s) of the chain.

Q 1. How do you envision search for similar documents, or search that finds documents?

Search for similar documents is not difficult, because we have a “query” document and we will be able to extract all search terms that we need ( via document identifier string): e.g. chain ids, substitute molecule(s) identifiers, etc. Will we need to provide ability for a human to search data store of SPL documents?

Q 2. I have examined SPL documents with “mixture component” moieties ( from [SDFs for srsutil](file:///C:\wiki\spaces\SD\pages\162463745\SDFs+for+srsutil) ). The structure of documents is straightforward: main substance contains collection of “mixture component“ moieties that define chemical structure via “subjectOf/characteristic[ code[ @code = 'C103240' ] ]“. Are there any other variants of mixture representation?