## NZOR Data Harvest and Import

### Harvesting

The main purpose of the harvesting module is to facilitate the process of gathering data from external data sources such as NZOR providers and synchronising that data with the provider data within NZOR.

|  |
| --- |
| Overview of Harvesting Module |
|  |

Main Functions:

* Scheduling of harvesting from OAI providers
* Validation of data prior to import
* Linking of provider data with system lookup tables
* Persistence data to the prov tables in the NZOR\_Data database

### Importing

Data for import must be represented using the NZOR Provider XML schema. <http://www.nzor.org.nz/schema/provider/101/nzor_provider.xsd>

Rules:

* When importing a provider record that already exists, the original provider record is update/overwritten with the new information
* Updates should be logged to an audit table
* If an update alters a Concept, the following needs to be considered:
  + Will this result in a concept being removed from a Consensus name – if so does this leave any hanging Names without a concept defining the parent (ie are there any names that are children of the name the concept was removed for, that do not have a provider concept themselves that defines the parent linkage – if so then on a refresh the name may be left with no parent).
  + Was the concept provided as a NameBasedConcept (ie there will be no ProviderRecordIDs for the Concepts and ConceptRelationships). If so, then the Concept and ConceptRelationship for the ConceptRelationshipType being updated will need to be removed and replaced with the new information – because this is a remove the above consideration will also need to be taken into account.
* If an update removes a consensus record, then checks need to be done to make sure there are no other consensus records relying on the removed consensus record.

### Concepts

There are several ways that Concepts can be provided by a provider – either as full **TaxonConcepts** or as **NameBasedConcepts**.

These two types of concepts roughly map to taxonomic concepts and nomenclatural concepts, where nomenclatural concepts are the name relationships as defined by the circumscription of the name, following nomenclatural codes, and taxonomic concepts are subjective relationships. For this reason you can have two “types” of Parent Name – one nomenclturally defined, the other a subjective placement of the name in the taxonomic hierarchy.

The full taxonomic concept model is as follows:

Taxon Name

Taxon Name

Taxon Concept

Relationship Type

Concept Relationship

Taxon Concept

Example:

Name ID = N1, Aus bus

Concept ID = C1, for N1

Concept Relationship, ID = CR1, C1 has parent C2

Concept ID = C2, for N2

Name ID = N2 Aus

Where all parts are required with IDs that the provider must maintain.

The NameBasedConcept has the following structure:

Name Based Concept

Taxon Name

(Accepted)

Taxon Name

(Parent)

Taxon Name

Name ID = N1, Aus bus

NBC, N1 has parent N2

Name ID = N2, Aus

Where only the Name IDs are maintained.

The problem comes when trying to import these name based concepts into the NZOR data (which has the structure of the first example).  We need to either put the NameBasedConcepts into the N-C-CR-C-N structure, generating temporary IDs or create more tables to handle NameBasedConcepts.  I prefer the former option as this means we only have one place to maintain Concept information.

On Import of Name Based Concepts:

* For the Name the Name based concept is provided for, remove all Concepts and Relationships of the same type (ie if ParentName is provided then remove the concept/relationship of that type, if preferred is provided remove that etc)
* Re-insert the Concepts and Relationship for the NameBasedConcept that is provided
* Leave the Provider Record ID for the Concepts and Concept Relationships as NULL – this indicates that these concepts were generated from a Name Based Concept.