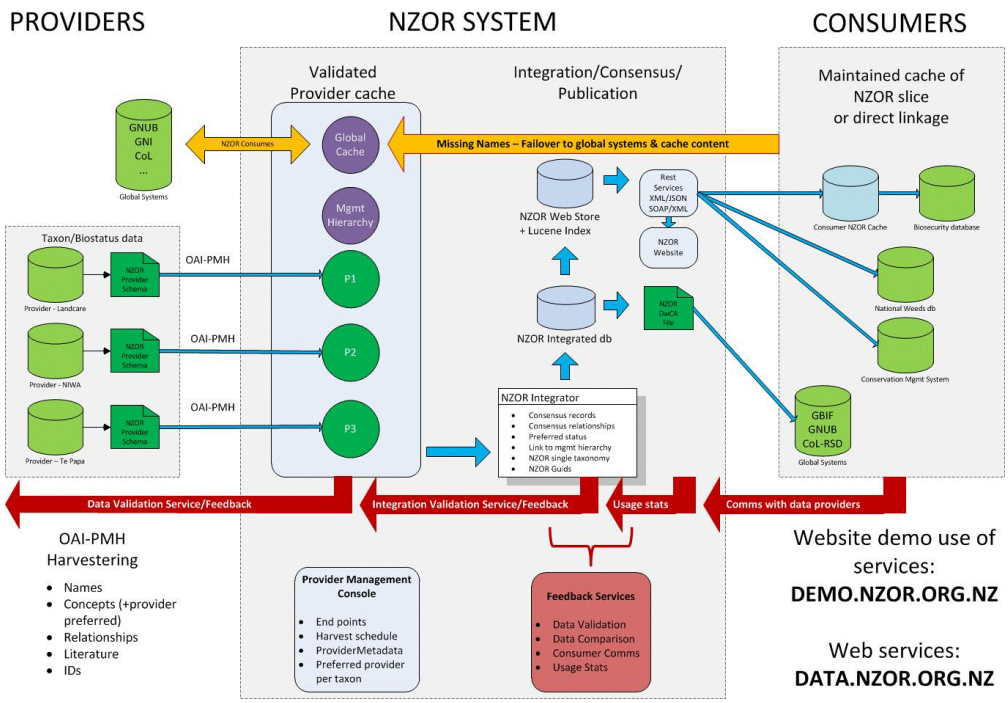
NZOR Overview

NZOR Technical Architecture

This section provides a semi-technical overview of the NZOR system.

The NZOR system has a number of linked components:

- A community of Data Providers who manage taxonomic data in their own database systems
- A data harvesting system managed by the NZOR system for regularly collecting data from data providers
- An integration engine which creates a single dataset from all data provider records and reconciles data which may overlap in content
- A data publishing system which transforms the single dataset into an optimised form for use by end users, which is exposed over the internet as a set of web-services
- An administration module for managing data provider content, data validation processes and a number of other activities
- A community of data consumers who use the NZOR web services to support local database systems, generally through the maintenance of a local cache of NZOR data



Harvesting

The NZOR system harvests taxonomic data from multiple data providers. Each data provider maintains a taxonomic database which may be in a variety of forms, data structures, and software platforms. Data elements within their system are mapped to a common data standard (the NZOR data provider schema). The NZOR Provider Schema is based on the TDWG Taxon Concept Schema Standard (TCS). The standard data elements are then made available for automatic harvesting over the internet using a generic harvesting protocol called the Open Archives Initiative for Metadata Harvesting (OAI-PMH). The combination of the mapping/harvesting interface is called a 'wrapper'.

The NZOR system regularly harvests provider data and imports it into a local repository. At the same time the harvester is carrying out a number of data validation and consistency checks. Eventually reports will provide feedback to data providers.

The core harvested data elements are the scientific and vernacular names of organisms. These are recorded as facts in the cited literature (published concepts). These published concepts are often associated with one or more asserted taxonomic relationships (preferred name, parent name). Data Providers indicate which of these taxon concepts they endorse (flagged as 'in use'). In addition information is mobilised on assertions relating to presence /absence of a taxon in New Zealand.

Integration

The combined set of harvested data is then integrated by the NZOR system into a single dataset. The integration process employs a decision tree to ensure that data elements from two or more providers which may overlap are dealt with appropriately. For example the same scientific name may come from three providers, perhaps a spelling error in one of them. The decision tree allows such differences to be detected and dealt with automatically. The result is a single NZOR dataset of integrated records with linked contributions from one or more data providers. In most cases the data mobilised by individual data providers does not collide during the integration process but the NZOR infrastructure is designed to be scalable to many data providers with large amounts of data and potentially substantial taxonomic overlap.

The integration process may be summarised:

1. Discover reconciliation groups of same names from all providers (equivalent name strings taking into account spelling variations)
2. Create/update simple majority consensus nomenclatural records linked to groups and create a persistent NZOR name GUID
3. Discover equivalent published concepts delivered by multiple providers
4. Create/update an NZOR concept records linked to a persistent NZOR concept GUID
5. Create simple majority concept relationship records for the preferred name and the parent name
6. Create/update NZOR single consensus taxonomic view from endorsed provider concepts
7. Break any deadlocks by following endorsed concepts from preferred provider for a defined taxonomic group
8. Track changes to NZOR taxonomy over time

Publication

The single NZOR dataset of integrated records requires transformation into a form that is optimised and indexed (Lucene) for querying. In addition Taxa Match algorithms are employed to parse and optimise the searching of organism names. End-users may submit queries on the dataset through a set of standard web-services. NZOR is designed to provide web-services in a RESTful format and SOAP. The result of a query conforms to the NZOR consumer schema and may be represented in a number of forms, e.g. XML, JSON.

NZOR Consumer Schema

NZOR provides data which conforms to standard form, the NZOR consumer schema. An NZOR dataset contains information on

1. metadata on the data providers
2. publications relevant to taxonomic names
3. scientific names and their nomenclatural details, and vernacular (common) names
4. taxonomic concepts, by which we mean the use of a name in a publication and the relationships asserted within publications about taxa, that one is a parent taxon of another, or is a synonym of another.
5. properties of a taxon, in particular the biostatus, by which we mean information on the presence/absence in a defined geographical region.