## NZOR Data Integration

Approaches:

1. Direct connection to database, get next record to integrate, match and update the records in the database
2. As in 1, but multithreaded (DB is the bottle neck)
3. Load all data for integration into memory, then process the data in memory first, then save all changes to the DB (multi-threaded)
4. As in 3, but run on a grid computing environment

Results:

1. 0.8 records per second
2. 1 record per second (threading not much advantage because DB is a bottleneck ??)
3. 17 records per second
4. ?

Conditions for generating another thread for the multi-threaded options include:

* Cannot integrate a name where the parent of that name has not been successfully integrated
* Cannot integrate a name that has any siblings that are currently being integrated

Possible improvements:

* Per thread, create 25 threads to process all provider names beginning with the same letter (as these are not likely to clash), or some other clustering approach – and perhaps only if there is > 100 child names ??
* Run a clustering algorithm over the provider names that need integrating before the integration is run to cluster unlikely matches (eg by first letter of name)

## Workflow

Importer

Provider

NZOR

Harvester

Multi-threaded, grid computing environment

Submitted CSV dataset

Integration Dataset

Results merged back into core NZOR data

Mapping to Integration Dataset Format

Integrator

Results returned to submitter

Integration Result Dataset