The National Library of Australia uses Voyager, an EXLibris product, as its local library management system. Bibliographic and holdings data is extracted from Voyager and shared or reused in other systems such as the local VuFind catalogue, in-house built Digital Collections Manager (DCM), Copyright Status Tool, Libraries Australia, Trove and OCLC Worldcat. This sharing of records and reuse of data exposes data quality issues that can impact on users’ access to our collection, services reliant on accurate data such as search limits and as a source of copy cataloguing. How do we identify data quality issues and correct them in the most efficient way?

We implemented Voyager in late 2003 and are currently running version 7.2.2. Voyager does not currently provide a great deal of functionality for making bulk changes to data so we have been forced to find a combination of other software solutions to help us manipulate and correct data from our large database; e.g. our database contains:

* 819,355 authority records
* 4,865,393 bib records
* 5,269,032 holding records
* 4,974,586 item records
* Plus other stuff

Tools we currently use to help us with the identification and correction of bad data (with a brief summary of what they are used for):

* MS Access software, which is used to identify data from the Voyager database (supported and recommended software by ExLibris).
* Voyager Webadmin tool, which allows the Voyager Systems Librarian to bulk export (or extract ) authority, bibliographic & holdings data in MARC format from the Voyager database.
* MarcEdit (free MARC editing utility from Terry Reece, Oregon State University), which is used to edit the extracted records. This utility provides an easy to use interface with many editing functions but also provides a more powerful option to use “regular expressions” to help match and define changes to data ( **Definition (if needed):** A regular expression is a set of pattern matching rules encoded in a string according to certain syntax rules. Although the syntax is somewhat complex it is very powerful and allows much more useful pattern matching than say simple wildcards like ? and \*.
	+ The Library’s Voyager system librarians are currently doing a remote course on perl programming language to further help with this process. We are in the very early stages of learning but are still very hopeful perl will help us to do more data changes more efficiently.
* The proprietary Voyager Webadmin tool can also be used to re-import in bulk mode the edited records back to the Voyager database. There are other non Voyager tools that we use to also do this and these tools can be used instead of MarcEdit for the editing of some data. Individual workflows and requirements rather than functionality can determine the best or most suitable option for Voyager libraries.
	+ These additional tools are unique to Voyager customers and have been developed by Gary Strawn from Northwestern University Library, Evanston Illinois. Gary has worked with the Voyager system for many years and he has developed a number of cataloguing utilities that can be run with Voyager. There has been a great demand for these types of tools as Voyager doesn’t include much of this functionality in the Voyager cataloguing module; as mentioned earlier. Some of Gary’s tools we use regularly are:
		- Authority delete
		- Bibliographic delete
		- Location changer (many holdings and item data fields)
		- Record reloader
	+ Our IT support staff have also been able to develop some supplementary tools for us to fill gaps in tasks or data not included in the aforementioned tools.
* Some of the data issues we are dealing with:
	+ **Historical data**; e.g. obsolete fields, indicators and subfields, including fixed field codes. The Library’s first local Library Management System, Dynix, was populated with bibliographic records and holdings data sourced from ABN. There were a number of odd ways ABN handled some specific data fields and we are still identifying and correcting these. At the time, as with it’s replacement Kinetica, staff catalogued on these systems and records with ANL holdings were pushed to Dynix via nightly file loads. Basically there were issues with this process especially when there were multiple holdings involved or an expectation that an edited record would match and overlay an existing record on Dynix. We are still identifying and fixing remnants of these processes.
	+ **Bulk import of large collection sets of bibliographic records;** e.g. microform and online databases). In many cases these records are based on print records and are system generated or manipulated by the record vendor. There can be issues with fixed field data reflecting the print or original format rather than the reproduction, spelling errors, unverified headings etc. all exacerbated by the fact some collections involve several hundreds of thousands of individual bib records.
	+ **Ongoing data quality issues.** Bad data can be created by individual staff on a daily basis, despite the best training. As well, poor data editing by system librarians trying to define bulk fixes does occur. The reality is, it is impossible for us to monitor all new and modified data in the database. Voyager provides some system defined validation checking of headings and of some coded data but not all data can be checked in this way. Any further checking requires some manual review although hopefully it can be corrected via more efficient and effective system batch changes.

We are currently targeting one category of errors. The Library has closed access collections so users need to request items via an electronic call slip system. This system is very efficient but relies on accurate data to determine which stack areas call slips are printed. We are currently monitoring more closely, with the help of Excel macros, inconsistent holding and item data that affects accessibility to collection items via electronic call slips. Although Excel macros help us to more easily identify inconsistent data, the data corrections themselves are very labour intensive and generally done manually.

* + **Updating and maintaining LC subject headings**; this work is well nigh impossible with just two staff in the team. We sometimes piggy back from the work Libraries Australia staff do and vice versa to try to keep up with MAJOR changes. We can’t keep up with all LC changes with the existing staff.

**In conclusion:**

Accurate and reliable data may seem like a pipedream, but we will keep on chipping away at trying to achieve the cleanest database possible; and try not to lose our sanity in the meantime.

We would love to hear from others who follow different practices with their data cleanup. I am sure we can all learn from sharing our knowledge about this popular topic that is dear to the hearts of systems librarians.