Metadata on the move

Off to Paris next week, but whoops! I forgot to renew my passport. No problem – I'll borrow my son's.
Of course, he is a young man, fair-haired and six feet tall, while I am female, darkish, shortish and oldish, so I'll have to make a few adjustments to the photo and to any other elements which relate to my personal description.
What could possibly go wrong?
Well, I might end up in gaol, which would prevent me from travelling at all. If the Border Agency happens to be cutting corners that day, I might avoid detection and get to France. But that could cause another set of problems. Records might then indicate that my son is in France, while I am still in the UK. If someone really needed to get hold of one of us, they might get the wrong person, or no one at all.
Because, even if I change the descriptive bits on the passport, it still has a serial number which is associated uniquely with my son, and so goes with *his* description, *his* attributes, *his* origins and travel history, not mine.

Nowadays catalogue records are travellers too. Many agencies continually import records from WorldCat, RLUK, the Library of Congress, the British Library and others, and export their own records back to WorldCat and RLUK – to say nothing of harvests and conversions from non-MARC metadata. And records, like human travellers, have identities and histories, which are enshrined in various serial numbers and other identifiers. These identifiers associate each record with a particular resource, and by implication with key aspects of its description, for instance its title, content, date of issue and physical format. Together the identifiers and descriptive elements act rather like a passport. The identifiers should make it possible for agencies to export their data without fear that it will be wrongly associated with data for other resources and to import data without fear that it will be of the wrong type or for the wrong resources.

For instance, the record below (p. 7), from OCLC's WorldCat, has the following identifiers (in red):

- OCLC system number (035); this will be preserved if the record is downloaded to another database
- National bibliography number (016); other records might have a BNB number (015) and/or a Library of Congress control number (010), doing similar jobs.
- Codes of the agencies which created or edited the record (040 $a, $c, $d), providing it with a sort of individual origins-and-travel history.

Just as for human travellers, if the descriptive elements within a record change radically they conflict with that record's identifiers, with the result that the record either fails to travel at all or is associated with the wrong resource and misleads people about where that resource may be found.

Nowadays records even have something like nationalities, the language-of-cataloguing-agency recorded in 040 $b, in this case French (in green). This is not the language of the resource, but the language used in the non-transcribed elements of the record, e.g. for physical description and notes. WorldCat now has records from many different language communities, and, to allow users to find records in their own language, has ruled that a change to the language-of-cataloguing requires a new record. That means that the identifying numbers in a record's 035 fields are associated not just with the resource catalogued but also with the language in which it is catalogued. National bibliography numbers may also imply a particular language of cataloguing.

```
LDR    cam a2200385Mi 4500
008    150428t20152015fr g b 001 0dfre d
016 7_ $a FRBNF443319120000005 $2FrPBN
020 __ $a9782200601508 (br)
020 __ $a2200601506 (br)
024 30 $a9782200601508
035 __ $a(OCLC)908129047
040 __ $aZWZ$bfre$encafnor$cZWZ$dBDF
082 04 $a306.8
100 1_ $aKaufmann, Jean-Claude$d(1948-....).$4aut
245 10 $aCasseroles, amour et crises :$bce que cuisiner veut dire /$cJean-Claude Kaufmann.
250 __ $a[Nouvelle édition].
260 __ $aParis :$bA. Colin,$cDL 2015, cop. 2015.
300    __ $a1 vol. (366 p.) :$b couv. ill. ;$c 22 cm.
336    __ $btxt$2rdacontent
337    __ $bn$2rdamedia
337    __ $bn$2isbdmedia
504    __ $aBibliogr. p. 347-358. Index.
650 _7 $aCuisine$x Aspect social$zFrance$y1990-.... $xEnquêtes.$2ram
650 _7 $aAlimentation$xAspect social$zFrance$y 1990-.... $xEnquêtes.$2ram
650 _7 $aFamille$xSociologie$zFrance$y1990-....$2ram
650 _7 $aSociologie du quotidien$zFrance$y1990-....$2ram
856 42 $3Notice et cote du catalogue de la Bibliothèque nationale de France$uhttp://catalogue.bnf.fr/
ark:/12148/cb44331912w
```

Has our cataloguing culture kept up?

Although I am employed by the Bodleian Libraries in Oxford, I work with all the cataloguers in Oxford University's Libraries Information System (OLIS), providing training, documentation, advice and a bit of quality control. Altogether I work with nearly 200 cataloguers in nearly 100 libraries scattered over several miles, half of which (college libraries and some departmental libraries) are completely independent of the Bodleian. I can't be looking over everyone's shoulders, so for quality control I have to depend on reports, when I know what to look out for, and otherwise on serendipity – just happening to come across records which indicate misunderstandings or knowledge gaps. So sometimes I get surprises.

All the *big* surprises in the last few years, the drop-everything-to-fight-this-fire surprises, have involved the kind of data which is used for *managing* records both internally and externally – the elements which control things like whether a record should be edited, or upgraded, or overwritten, or exported to COPAC or WorldCat and, particularly, how our exported records will be handled by the receiving agency. Fires of this kind have proved particularly difficult to fight – smouldering unnoticed for years, taking hold in several locations at once and with a trick of breaking out afresh when we hope they have been damped down.

The two biggest surprises related to 'passport control', and I would like to share what we learned from them, because they indicated that our cataloguing culture had not kept up with the changing cataloguing environment. They showed that our cataloguers found it much easier to appreciate the importance of good descriptive data and access points than the importance of the kinds of codes and identifiers which are primarily intended to be read by machines for automated processes, especially when the processes involve databases other than our own. But this kind of data is absolutely key to our ability to meet the growing challenges of dwindling resources and the growing opportunities of linked, machine-actionable data. We know that these problems are not confined to the Bodleian. Some of our faulty records had been copied from other databases with the problems already present – from OCLC or RLUK, even a few from the Library of Congress and the British National Bibliography.

I am not at all intending to criticise my own colleagues or cataloguers in other agencies. When I was first trained to catalogue, just over a decade ago, I did not need to know very much beyond AACR2, MARC21, LCSH, our local software, and some fairly basic information about the external databases from which we could copy records. Now our cataloguers are expected to understand and make best use of a vast, varied and ever-changing network of data and processes, both local and external. They have to be up to date with what to trust, what to check, what to adapt, what to avoid entirely, so that they neither waste time on unnecessary checking of good data nor accept data which is not fit for purpose. It's a big ask. Senior staff may design and document 'efficient' new streamlined workflows and algorithms, but that is not enough – the workflows and algorithms will not work efficiently until we win the hearts and minds of all our colleagues to care about *all* the kinds of data which the workflows and algorithms use.

Records from foreign-language agencies

The first shock hit us in the summer of 2012, just when we were focusing on our preparations for RDA implementation and really, really did not want to have to think about anything else. We noticed a record which did not use Unicode and so displayed oddly in our system.

24510	|a ❖Die❖ geretteten Götter aus dem Palast vom Tell Halaf |b Begleitbuch zur Sonderausstellung des Vorderasiatischen Museums "Die geretteten Götter aus dem Palast vom Tell Halaf" vom 28.1. - 14.8.2011 im Pergamonmuseum |c für das Vorderasiatische Museum - Staaliche Museen zu Berlin hrsg. von Nadja Cholidis ...

Its 040 $b and $e showed that it was created by a German-language agency to a German standard (Regelwerk für Alphabetische Katalogisierung Wissenschaftliche Bibliotheken), and in other respects it was unsuitable for our use. For instance, the physical description was in German and there were no English-language subject headings.

But that was not the shock.

Initially we thought that the record was more or less a one-off, and that all we needed to do was to remind our cataloguers that OCLC included non-English records, and make sure that everyone knew what to look out for and avoid. The shock was when we were told by several managers that their cataloguers could not get through their workload without downloading non-English records, *and had been downloading them for years*. In fact, there were already several thousand records in OLIS coded as non-English. We just hadn't noticed them before, among our millions of records. Most had been more or less fully translated and adapted to English-language standards, but were still coded as French, German, Dutch, Spanish, Italian, while a few had not been translated or adapted at all, and so still had physical descriptions and notes in languages other than English.

Because our attention had been absorbed by RDA and, before that, a system migration, we had not given much thought in the previous few years to the fast-increasing availability of foreign-agency records, and many of our cataloguers were actually unaware of the significance of the code in 040 $b. It had hardly occurred to me or my immediate colleagues that anyone might be tempted to use a record which had funny 300 fields and used unfamiliar standards and authorities and subject headings – surely it would be simpler to catalogue originally? That was a failure on my own part to engage with this new source of potentially valuable data, because when original cataloguing of a foreign-language resource would involve lots of diacritics or nonroman scripts and transliteration, adapting a foreign-language record can be an attractive option.

It is obvious that untranslated or semi-translated records are not fit for purpose. Users might well be perplexed by finding in an English-language catalogue a physical description such as '24 cm met CD' ('met' is Dutch for 'with') or a note such as 'Lizenzpflichtig' (German for 'Subject to license').

And it is not just the language itself which might cause problems for users: foreign-language records often use different cataloguing rules, different versions of MARC, different name authorities and different subject systems, so a great deal of editing may be necessary to make sure that users can find the resource in the usual way and discover its relationships to other resources in the collection. But if a record has been thoroughly translated into English and edited to our usual standards, does it really matter if it still has in 040 $b a code which claims that it is in a language other than English and/or has a number in 035 which identifies it with a record in a language other than English?

It would not matter if the record was never going to travel anywhere else ever again; but in fact all our finalised full-level records (and many less than full) are exported to OCLC and RLUK, and so it matters in the following ways:

- OCLC will often reject entirely a record which has an 035 number identifying it with an existing OCLC record but differs from that record in key elements such as language of cataloguing or extent, so changing 'ger' to 'eng' in 040 $b or changing 'Seite' to 'pages' in 300 $a without giving the record a new identity might mean that the record does not reach WorldCat, and WorldCat users will not know that we have a copy.
- If translated records with unchanged language codes do slip into OCLC, they will be a snare and a delusion. For instance, a Spanish-language agency which wants to harvest Spanish-language records, selected by language of cataloguing (040$b=spa), might actually get a proportion of records which have been translated into English. This is particularly disturbing because it feels like a breach of trust and a betrayal of our common values and objectives. If agencies cannot be reasonably confident that data which is labelled as being of particular types really is of those types, all our initiatives to maintain and improve services by exchanging and linking data will be undermined.

It's much the same as if I tried to travel on my son's passport: either I wouldn't get to travel at all, or I would slip through the checks and then people who wanted to find my son might get me instead.

We had to get ourselves sorted fast:

- For the thousands of existing OLIS records with non-English 040 $b, we could only do a rough global correction, because there was so much variety as to how far they had been checked and/or translated; so, sadly, all those records had to be downgraded.
- A systems colleague designed a brilliant fix to make it possible to adapt downloaded foreign-language records safely. It recodes the records correctly as new English-language records, removes the history,[1] and adds a local field with a warning that the record must be checked thoroughly for conformity with MARC21, RDA, LCSH, etc. Cataloguers are not supposed to delete this field until the checks are completed, and meanwhile it prevents export.
- I added a new check to our record-checking software (Marc Report) to warn cataloguers if they saved a record as full-level and ready for export while it still had a code other than 'eng' in 040 $b.

---

[1] Some delegates at the CIG 16 conference were unhappy about removing the history of work done by other agencies, and it certainly does not feel good to remove an indication of intellectual input; but a new record by definition has no history of its own, and it is not usual to record 'ancestry' for anything less than a whole record. A cataloguer who re-purposes some of the data from an existing record takes over responsibility for that data, just as s/he would if s/he created an original record but saved time by copying over data from some other record field-by-field. OLIS does actually keep some local data which indicates the record's previous incarnation, and Alan Danskin of the British Library mentioned that ancestry could be recorded more formally by using MARC field 038, Record Content Licensor, but this is not common practice. Perhaps this is an area for more discussion among the cataloguing community. The situation has crept up on us without much opportunity to think through all the implications.

But cleaning up the bad data and making safe procedures available was only the beginning. We still needed to make our cataloguers fully aware of the importance of using the fix and making the checks, and that has required a very serious and sustained effort: not just documentation, automated warnings and general reminders, but in many cases person-to-person explanations, talking through how users and staff, both in OLIS and in other agencies, might be misled or inconvenienced, so that the issue no longer seemed like a mere technicality. This 'humanising' of record-management data is particularly difficult for us because our cataloguers are so scattered, but it is very worthwhile.

After four years things are much better, but we still occasionally come across records where a cataloguer has made the warnings go away by making little edits to the elements which trigger warning messages, for instance by deleting 040 $b, rather than using the proper fix and procedures to assign a new identity.

'Unresolved' records

After our last system migration it was not possible to set up a regular export to OCLC for quite some time. When we eventually exported the backlog, several thousand records were rejected. Many of these were translated records which had retained the 040 $b or 035 of the original, as explained above, while others were technically incorrect, such as records for mathematical or scientific resources with untransliterated Greek characters in the title fields. But very many were rejected simply because they had been derived from OCLC and retained the 035 fields of the master records from which they were derived, but they had been edited in ways that made them seem to be for different resources from those covered by the master records, with the result that OCLC's deduplication algorithm could not process them.

In a minority of cases the edited records really were for exactly the same resource as the original OCLC record but had been improved out of recognition (e.g. brief OCLC records enhanced to antiquarian standards), and in such cases the only solution is to improve the OCLC record correspondingly. Far more often the records really had been edited to match different resources, for instance by changing the edition statement or ISBN or publication date or carrier or format or by extending a single-part record to cover a multipart resource.

Like the foreign-agency records, these illegitimately adapted records cause two types of problem:

- If our records are rejected, WorldCat users will not discover that we have those resources.
- Some adapted records do get through, cluster with the wrong master records, and give users the impression that we have a resource which we do not have, wasting their time and disappointing them.

As with the foreign-agency records, what was really startling was that records were being illegitimately adapted not just occasionally by oversight, but quite routinely by experienced and careful cataloguers. In some cases a single OCLC record had been downloaded as many as four times for separate editions of a work, with the result that we had four different OLIS records all with the same OCLC identifier in the 035 field.

On the next page is an example, showing key fields in an OCLC record and in an OLIS record which was derived from it and carefully enhanced to antiquarian standards. The OCLC identifier in field 035 is retained in the OLIS record, but the title proper, (lack of) edition statement, place, publisher, date and physical description are all very different, showing that it is being used for a quite different resource. In effect, it is using another record's passport. When it tried to travel to OCLC it was caught and sent back.

OCLC record

035 __ $a(OCLC) 19508660
100 1_ $aVarennes, Claude de.
245 13 $aLe voyage de France, dressé pour l'instruction & commodité tant des françois que des estrangers.
250 __ $a4. ed., corr. & augm.
260 __ $aRoven, $bJ. Caillové,$c1647.
300 __ $a10 preliminary leaves, 60, 304, [29] pages

OLIS record

035 __ $a(OCLC) 19508660
100 1_ $aVarennes, Claude de.
245 13 $aLe voyage de France, dressé pour l'instruction & commodité des François & Estrangers.:$bAvec une description des Chemins, pour aller & venir par tout le Monde. Tres-necessaire aux Voyageurs. /$cCorrigé & augmenté par le Sieur Du Verdier, Historiographe du Roy..
260 __ $aA Paris,:$bChez Michel Bobin, au troisiesme Pillier de la Grand'Salle du Palais, à l'Esperance.,$c1655..
300 __ $a8 unnumbered pages, folded map, 492, 4 unnumbered pages :$billustrations ;$c8°.

The solution for this problem was in principle quite simple. No new processes or procedures were needed, and the existing documentation described the proper processes and procedures correctly.[2] What was lacking was (i) a clear understanding that copying a whole record, identifiers and all, is not at all the same thing as just copying a lot of useful bits of data, and (ii) an appreciation of the human cost of using the wrong processes – the frustration and wasted effort for users, and the deflection of precious staff resources into laborious record-by-record analysis and repair. This particular hearts-and-minds campaign has been running for less than a year. There has been a great improvement, but some cataloguers do still sometimes forget, particularly generalists who catalogue only occasionally.

These are just two examples of the human cost of paying too little attention to the parts of a record which are intended to facilitate processing by machines. The importance of such elements can only increase with the challenges of dwindling resources and the opportunities of linked, machine-actionable data. If we are serious about sharing and caring, we nowadays have to be prepared to share and care on a scale so large that it can only be managed by mechanised processes. By exchanging data on a really large scale we keep our operations viable, enrich our services and support other information providers.

So cataloguers must learn to care about how they talk to machines. And for that to happen, people in jobs like mine must care about how they talk to cataloguers and remember to present mechanised procedures and processes in the context of the human needs and aspirations which they serve.

---

2. A cataloguer who wants to improve an existing OLIS record with data from a better record has to copy over that data field by field, but that is less tedious than it sounds, because our Aleph system allows multiple adjacent fields to be copied simultaneously. The important thing is just to avoid using the 'Duplicate Record' command.