Metadata Quality Control

Metadata – why does it matter?

The National Information Standards Organization (NISO) defines metadata as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource".[1] A common example would be a library catalogue record which lists the title, publisher, size, ISBN, keywords and shelf number of a particular book. But metadata also includes hidden preservation, administrative and technical information such as when a particular record was created and by whom, or what intellectual property rights are attached to an object.

Given the rise of new web technologies, an expanding number of encoding and descriptive standards, increased user expectations and the rapid development of digital repositories and cross-institutional funded projects, metadata is now a subject of much attention for its importance in facilitating access to digital resources, both within and across organisations. Metadata *quality*, therefore, is a matter that every institution must and should address on a consistent and regular basis.

What constitutes poor metadata?

"Good cataloguing is the foundation stone of librarianship. If you have an item and can't find it, you don't really have it".[2]

Common metadata errors include misspellings and other typographical errors, incomplete fields, empty fields, incorrect use of punctuation or keywords, duplication of records, inconsistent formatting of dates and legacy issues from previous cataloguing methods or standards. The result is that, for example, if the author 'Judy Blume' is entered manually as 'Jdy Blume' at the point of metadata creation, then that particular record will never be found when a user types the full name into the search field of the catalogue.

Ultimately, poor metadata renders content virtually un-findable, having detrimental knock-on affects for research, commercialisation, publication and customer services.

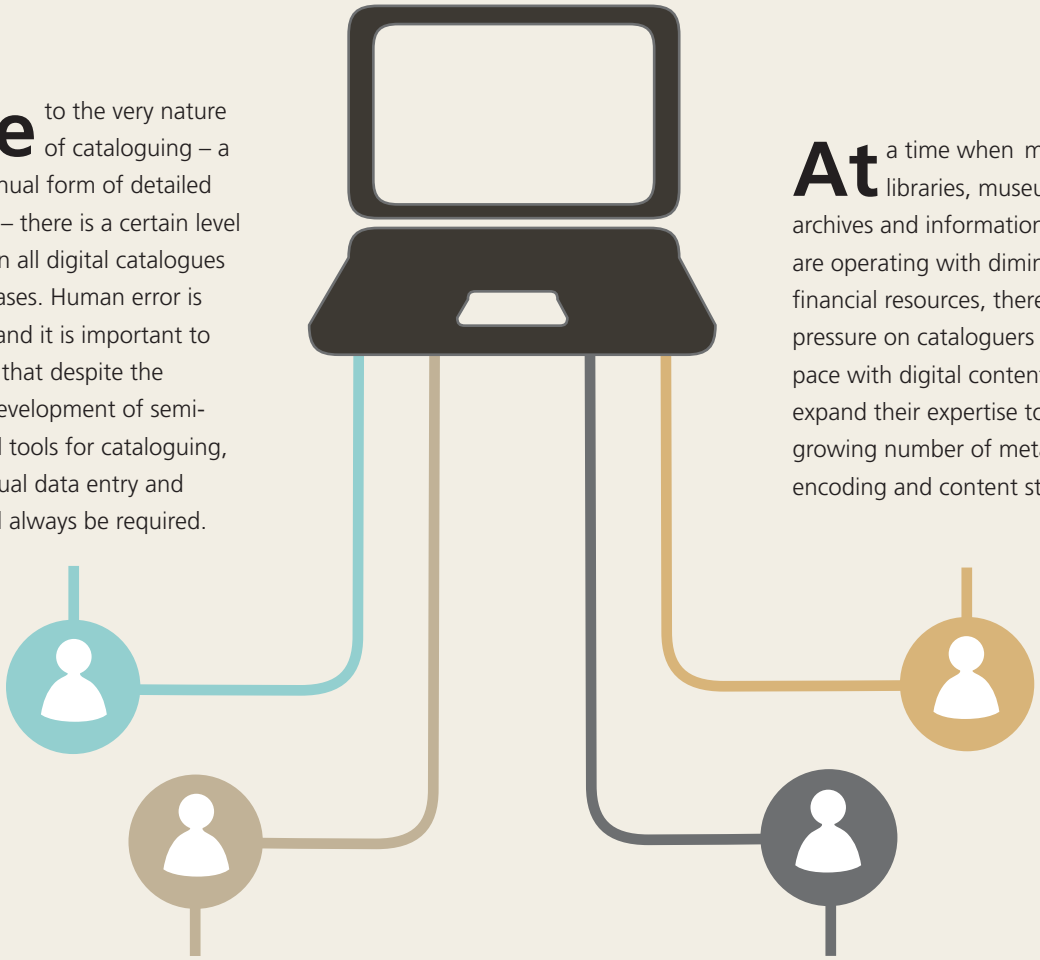
---

[1] http://www.niso.org/publications/press/UnderstandingMetadata.pdf
[2] Overholt, J. [John_Overholt]. (2014, July 29). Good cataloguing is the foundation stone of librarianship. If you have an item and can't find it, you don't really have it. [Tweet]. Retrieved from https://twitter.com/john_overholt/status/494118064863334400

Why does this happen?

Numerous challenges exist in the current information environment that impact on good quality metadata creation:

Due to the very nature of cataloguing – a mostly manual form of detailed data entry – there is a certain level of 'noise' in all digital catalogues and databases. Human error is inevitable and it is important to remember that despite the ongoing development of semi-automated tools for cataloguing, some manual data entry and editing will always be required.



At a time when many libraries, museums, archives and information centres are operating with diminished financial resources, there is more pressure on cataloguers to keep pace with digital content and to expand their expertise to include a growing number of metadata encoding and content standards.

At the same time there are more instances of non-professional metadata creation – crowd-sourcing to tag online images, content authors uploading their data to repositories, etc.

In the case where a library or archive forms a department within a larger organisation, it is sometimes difficult to convey the importance of the work, and metadata creation can be viewed by others as a tedious or irrelevant task. This leads to fewer opportunities for current staff to upskill as well as the danger that cataloguing issues will be ignored until they become a far more complex problem.

Nobody's perfect.

It is of course, important to note that *quality metadata is what is fit for purpose*, and what fulfils the functional requirements of your content or library management system. There is no perfect solution to metadata creation and regardless of developments in technology there will never be a time when all of your organisation's records are complete and clean. Every institution or organisation, whether a television broadcast archive or a national library, needs to utilise standards and vocabularies in a way that serves their own users, be that internal (staff members) or external. What is vital, however, is that metadata quality control is built into regular workflows at as early a stage as possible.

So what can you do?

Many commercial companies now have formal quality control practices built into their workflows. You can find many examples of these online, and they can be adapted to suit your own institution.

For libraries and archives, DRI recommends a **Five-Point Strategy:**

1 Best practice

Spend some time researching **best practice** for your workplace, whether that is an academic repository, a public library or a broadcast archive. There is a vast amount of information available online through official websites and through the information community on blogs and social media. The National Information Standard Organisation has an excellent set of broad metadata principles that can be used as the foundation of your quality control strategy.[3]

2 Clear cataloguing procedures

Have **clear cataloguing procedures**, especially if non-professional cataloguers are going to be responsible for creating metadata. Develop up to date guidelines that are readily available in hard and/or soft copy to your cataloguers. Make sure your interface is intuitively designed, with a clear-cut set of fields to complete. Consider implementing drop-down lists to control descriptive content. If not then perhaps you can input links to relevant indexes, thesauri or authority lists.

3 Training

Facilitate **training** for staff when required. Take advantage of bursaries and scholarships to attend training days in Ireland or the UK. Look for free webinars. DRI also runs regular seminars and training events at little or no cost. Remember that at a time when it is difficult to replace or hire new staff, upskilling and professional development must be granted the support required to maintain quality services.

4 Usability tests

Conduct **usability tests** with those who most regularly need to search and use your metadata. This may be members of your staff, members of the public or both. Short online surveys can collate a great deal of information about the requirements of those who are regularly using your databases, drawing attention to inconsistencies or difficulties they have. For internal users in particular, such as staff members who use your metadata to research for their own work, consider implementing a means by which they can report any mistakes that they find.

5 Regular quality audits

Organise consistent and **regular quality audits** of recent metadata samples. These reviews are especially important if changes have recently occurred in guidelines or workflows, so that you can track whether or not new instructions are being implemented by your cataloguers.

---

[3] http://www.niso.org/publications/rp/framework3.pdf

How to do a metadata quality audit

DRI recommends taking the following steps to conduct regular metadata quality assessments:

- Designate one or a small team of information professionals to take responsibility for the audit.
- Decide to what extent any mistakes found during the audit will be fixed within the live database.
- On a quarterly or biannual basis, upload a sample set of records to the software application OpenRefine.
- Use the Faceting and Cluster tools in OpenRefine to identify and record errors, such as misspellings, inconsistent use of capitalisation or blank cells.
- Compile the documentation so that any changes in quality can be noted over a period of time. This will be particularly useful if the organisation has recently started using new cataloguing methods.

Using Open Refine

Background

OpenRefine is a free, open source data wrangling tool. It began as Freebase Gridworks developed by Metaweb, and when Metaweb was acquired by Google in 2010, was rebranded as Google Refine. Since 2012 Google ceased active support of the tool, and it was renamed OpenRefine. The tool is a desktop based application that can be downloaded and then used locally. A subset of metadata can then be exported from your content management system and uploaded to OpenRefine as a new project.

It is important to note that the tool is not designed as a means of cleaning so that data can be re-imported back into the organisation's live database. The cleaning functions, powerful as they are, are used to prepare datasets for reuse, whether that is to reconcile with controlled vocabularies, perform named-entity extraction or publish as open Linked Data. What is valuable for the purposes of this document is the ease with which a dataset can be explored and examined, and the facility to quantify occurrences of misspellings, blank cells, punctuation and other errors.

Faceting

The OpenRefine interface looks like a spreadsheet – with rows of data running horizontally and columns running vertically – but it operates more like a database. Faceting can be used on each column to list all the different types of cell values and the number of times that those values occur. So a facet created in your Keywords column, for example, might show 20 entries for 'Weather' and three entries for 'Weathre'. You can then easily identify the three records with the misspellings.

Cluster and Edit

Applying the Cluster & Edit tool on one column will cluster words with close lexical matches, thereby identifying misspellings or inconsistent capitalisation. What is particularly useful, as with the Facet tool, is that the number of times that cell values are spelled in a unique way is recorded beside that value. Again, this can be recorded, and the records with the inconsistent spellings easily brought up and explored.

DOI: 10.3318/DRI.2015.1