NBK Draft Record Quality Measures

June 2020

1. Introduction

One aim of the NBK has been to support libraries in reviewing catalogue data and helping to raise quality where this may be needed, improving catalogues locally as well as for record share through Library Hub Cataloguing. Improved data quality also helps with record deduplication, benefitting both resource discovery and collection management. Now that the NBK is established we are beginning to plan what an 'Analytics' service might look like, to support NBK contributors with data review and upgrade.

For example:

- You may have concerns about the records you are receiving from a particular ebook supplier. An Analytics service might allow you to review those records to assess whether there is actually a problem and what the issues are to focus on in discussion with the supplier.
- You might wish to identify basic records from a past retrocon that need upgrading. We could then look at options to request 'better' records to either replace the record, or possibly to request particular fields that are missing from your records, thereby avoiding the loss of local content.

One important element of this work is an understanding of record quality and what a 'better' record might look like. Of course, there isn't necessarily a single answer and different libraries might have varying views on how they would like to see records ranked. So we also want to see how we could combine a measure of quality with other preferences.

Aims

The quality measures are not intended to be about the 'perfect' record, rather they are a practical working tool that will help us to support you in the review and upgrade of records where this is needed. The Analytics service will enable you to review records using standard measures such as missing or incomplete fields, but we also feel that a general measure of the quality of each record is valuable for two main reasons:

- To help you identify problem records by multiple criteria rather than just looking at them from the perspective of a single problem
- To give us the potential to identify records that are 'better' than your existing record, that we might then make available to you

The diversity of the records in the NBK, and the current usage constraints around some records, mean that we may not have a 'perfect' record for any document. But we do want to be able to offer the best available record(s) to support catalogue development. We are at the early stages of this work and there are a lot of questions around the best approaches to some of this activity which we will be working through as we begin to establish a trial version for exploring options with you over coming months. We will be learning as we go along and will be aiming to create something we can build on. However, the

record quality measures will be a foundational element of a new database to support analytics activity, so we want to get an initial version of these measures in place at the start.

Business case

As we think about how an Analytics service will develop, we will also be examining the business case to support the work involved in creating and supporting a new service. It may be that some aspects of such a service are chargeable, as an add-on to Jisc members and others, in order to enable us to proceed with this work.

Acknowledgements

We would like to thank all those who have provided us with feedback on this, and related, questions over the last couple of years. We have tried to incorporate comments we have received into the thinking underlying this work.

Associated documents

Data samples

The 'Sample record sets for NBK Draft Quality Measures' document provides some sample data sets covering a range of material types. Each data set gives a variety of records for the same document that have come from a range of libraries. You may find these useful as examples of data variation as you consider the quality measures and the likely effect of any changes you might wish to propose. But they are not intended to constrain you from looking more widely for your own examples.

Feedback questionnaire

We have provided a set of questions to consider whilst reviewing the draft approach to quality measures.

Responding to the consultation

If you have questions about any aspect of the proposed quality measures please contact us through the NBK mailbox **nbk@jisc.ac.uk**

We would like feedback on the quality measures, including completed questionnaires, by **Tuesday 30th June**

Please send your feedback to the NBK mailbox: **nbk@jisc.ac.uk**

2. NBK Quality Measures: Background

Criteria

Several criteria have guided our approach to the record quality measures:

- Transparency: We want the quality measures to be clear and straightforward so that if you are using the service you can understand why any record has been given a particular score. This makes it much easier for you the assess whether it is working as you might wish or whether there are issues we might want to address to improve the service.
- Sustainability: We want the quality measures to be sufficiently straightforward that they can evolve over time, in the light of feedback, or in response to changes in the underlying data and the data standards. Changes will require a full rebuild of the data so it would not be a regular event, but it is also not fixed. We wish to avoid complex algorithms that are only fully understood by the original developer and which are problematic to maintain because of the difficulty of understanding the implications of any change.
- Practicality: Since January 2020 we have been loading and updating between 1.25 and 8.25 million records a week, a total that is likely to grow with the number of NBK contributors. This means we need a process that can apply a quality measure to all the new and amended records in each update without impacting on the overall speed of processing.
- Flexibility: It will be possible to use the quality measures alone to work with the data, but we also want these to be options that in future might be combined with other preferences. For example:

    - you might want the best available RDA record for a document
    - you might want to review your map records to see where you have records of reasonable quality, but that are still missing fields you would like included, eg. a 255.
    - you might prefer a 'better' record from a particular source, as long as that record is above a certain quality threshold

We are looking at ways of measuring record quality that will give a straightforward and, hopefully, usable approach to supporting your work in record review and upgrade.

General Model

We want the approach to be flexible to allow different approaches and support different views of what a good record looks like. We have taken a two level approach:

- Level 1 - Breadth: The first check will assign an overall quality value to a record. The focus here is on the breadth of coverage, essentially the range of fields included. Fields are grouped into categories and where a record has any one of the fields in a category it is given a score of one. Adding the category scores gives you the level 1 quality score for that record. There are 17 categories, so this will give you a score range of 0-17. (See Section 3 for details of these categories)

    This can result in multiple records for the same document with the same score. So a second quality value helps to distinguish between these.

- Level 2 - Depth: In this second check the focus is on the depth of coverage, for example, rather than just checking that there is a 6XX field, looking at how many 6XX fields are included.

    The scale here can be as wide as the records are extensive.

- If both level 1 and level 2 checks come out with the same values for two records they will be regarded as being of equal quality.

The two-level approach offers flexibility, so the measures can be used independently depending on the data you are looking at and what you want from any 'better' records.

The grouping of fields into categories at level 1 generally brings together fields with similar functions, aiming to provide an overview of a record across the basic features required for effective local data use, including resource discovery. Assigning a field to a category is essentially giving it a weighting, so there is a balance to be struck between separation and merger of fields within categories. For example, bringing together classmark fields (050, 082) obviously reduces the detail, but to separate them out would give each type of classmark an equivalent weighting to the presence of an author in the record.

The separation of the more general fields (eg. 300) from more material specific fields (eg. 306) helps to provide greater granularity in the scoring. For example, for an audio-visual item it might help distinguish between an adequate but general record from one created by a specialist with more material specific details.

The full count of fields at level 2 gives the more detailed overview of the scale of the record. Two records with the same level 1 score could be quite different in terms of comprehensiveness, for example one record might include multiple 700 fields giving detailed contents for a compilation. But it is not without its limitations. A record with rather basic bibliographic content could theoretically get a high quality score as a result of a large number of subject headings, which might not be what you want.

The advantages and disadvantages of each approach are likely to become clearer with use. And it may well be that different libraries develop a preferred approach, or that the approach may vary depending on the nature of the data in question and the review activity. The intention is that the quality score can be combined with other options to give a flexible approach to record review.

The proposed model does not represent a final and fixed process. It will be dependent on development constraints and the effect of applying the tests in a working environment. It will also be subject to review and has been designed with consideration of the need to respond to feedback as well as changes to incoming data and data standards. However, after the initial development period we would not anticipate regular changes. Stability seems important to avoid potential confusion over the way the measures are applied, whilst any changes will require significant data changes so would take time to implement.

Data coverage

We will be creating a new database to support the Analytics service trial. A number of catalogues will be excluded, as the analytics seem unlikely to be of relevance to these libraries:

- Catalogues will be excluded where they are supplied in non-MARC formats
- Catalogues will be excluded where they have been converted into MARC from a non-MARC format before being supplied to us, and require pre-processing before we can load them

Otherwise, all data will be included in the new database in the form supplied, so that you have a full view of your own catalogue records.

Dropped records

The database will include records that have been dropped during the load and thus excluded from Library Hub Discover and Compare. This will give you a way of reviewing records dropped for any of several reasons:

- Records matching library-defined criteria for exclusion eg. a record with a specified location code and no other physical or electronic location
- Any record with no 245 is rejected and dropped from the load into Discover and Compare
- Any record that cannot be parsed is rejected and dropped from the load into Discover and Compare, eg. where the Leader is the incorrect length
- Where records have duplicate record IDs the second and subsequent records with the duplicate ID will be rejected and dropped from the load into Discover and Compare

Restrictions on record availability

However, where we look in future at the question of supplying 'better' records, the same data restrictions will apply as is the case for Library Hub Cataloguing:

- We will be working within the current licensing restrictions that relate to some records. A record with licensing restrictions will be given a quality score so the information is available to you as the originating library, but that record will not be made available to other libraries where there are restrictions on re-use of that data
- Where you have asked us not to make your catalogue available for shared cataloguing these records will be included in the analytics service, and will be given a quality score, but they will not be available to other libraries as part of a future record upgrade option

It is likely that to support both catalogue review and record upgrade we will need two parallel sets of quality scores:

- For record review all elements of the records would be included, and quality scores assigned, so you get a full picture of your catalogue
- However, for us to provide 'better' records in future we would want a second set of quality scores:
    - OCLC provenance records would be assigned a quality score that is based only on those fields that OCLC permit to be included in their 'mini-MARC' record.
    - we would exclude from the quality score those fields with a $5 which are unlikely to reflect the potential value of a record to another library.

Error checking

We will be checking for the presence of particular fields, but beyond this we want to know that the fields are accurate, in as far as that is possible. We wish to avoid giving a high 'quality' score to a record that has a good range of fields, but where the structure of those fields is poor, which might, in turn, reflect limitations in the record content.

To begin with the focus is on the things that are straightforward to check, as outlined below, so the MARC structure and clearly defined contents, such as the validity of codes at each character positions in the Leader. We are essentially looking at whether a field is 'well formed'. However, whilst we are checking the structural accuracy of the record, we are not attempting to check the 'relevance' of the content this stage. For example, we will check that the Leader character position 06 'c' is valid for that position, but we will not try to determine whether the record actually represents notated music.

Once the basics have been established, we are interested in expanding the data checking to include some basic content review. So whether a map has the correct basic codes in the Leader and 007; or whether the non-filing characters are correct for English language materials. However, the further we move into checking record content the greater the level of uncertainty this may introduce, for example, trying to distinguish between a medical or geographic atlas can be problematic in a brief record with Leader cp06 a, whilst a 245$a starting 'A is for Apple…' might not want non-filing characters. So, these

types of check might be considered in future for data review rather than necessarily informing the record quality measures.

Proposed error checks

We will need to experiment with processing speeds, but the aim is that if there are errors in the field it will not be included in the quality score. We will be looking at applying the following checks:

Leader

- Are the codes in the Leader correct for each character position?

007

- Do the first two character positions of the 007 contain valid codes?

008

- Is the 008 field the correct length?
- Does the 008 field have content other than space or |?
- Are the codes in the 008 correct for each character position?

All other fields

- Does the field have content other than space or punctuation?
- Is the field content all uppercase?
- Does a field have valid indicators?
- Are the subfield codes valid?
- Are any subfield repeats correct?
- Are any field repeats correct?

The above checks do not assert that the field content is correct in the context of the material type or the individual document. So a 100 field may be correctly structured but may contain the wrong author for the work described by the record.

In addition to general checks we would like to look at filtering out less ‘standard’ or less complete fields, so we have suggested some field specific checks as indicated in the ‘table of fields’ below. These are intended to try and exclude fields where the content is largely of local relevance to the originating library, or the content is very limited eg:

- A 650 containing local subject terms
- A 780 that only contains a $w or standard number

3. Draft Quality Measures

Field coverage

We have taken as a starting point the BIBCO Standard record (https://www.loc.gov/aba/pcc/bibco/documents/PCC-RDA-BSR.pdf) and CONSER Standard Record (https://www.loc.gov/aba/pcc/conser/documents/CONSER-RDA-CSR.pdf). Including fields to provide for both basic record content as well as those of particular relevance to specific materials.

The table below shows the fields suggested for inclusion in the record quality scores, grouped into categories. Each category counts as '1' in generating the Level 1 score for a document, whilst the fields are counted individually, with repeats, for the level 2 score. So a record with all three of a good Leader, 007 and 008 would get a level 1 score of '1', and a level 2 score of '3'.

Table of fields for generating record quality scores

Category	Field(s)
Leader & Control fields	Includes • Leader • 007 • 008 If there is no 008, or it is empty, the record can only get a maximum score of 1
Identifier	One or more of the following: • 020 - International Standard Book Number. Must have $a or $z • 022 - International Standard Serial Number • 024 - Other Standard Identifier. First indicator 0,1,2,3,4,7 • 026 - Fingerprint Identifier • 027 - Standard Technical Report Number • 028 - Publisher or Distributor Number • 030 - CODEN Designation • 088 – Report Number
Authentication Code	An 042 that includes one of the following codes: • lc - Library of Congress • lccopycat - LC Copy Cataloging • lcd - CONSER full authority application • lcnuc - National Union Catalog • nsdp - National Serials Data Program • pcc - Program for Cooperative Cataloging • ukblcatcopy - British Library derived cataloging • ukblderived - British Library derived cataloging

	• ukblsr - British Library Standard Record • ukscp - UK Legal Deposit Libraries' Shared Cataloguing Programme Record
Coded language/place/time	Includes one of the following: • 041 Language • 043 Geographic area code • 045 Time Period of Content • 046 Special Coded Dates
Class number	One or more of the following: • 050 - Library of Congress Call Number • 082 - Dewey Decimal Classification Number
Main entry	One or more of the following: • 100 - Main Entry - Personal Name • 110 - Main Entry - Corporate Name • 111 - Main Entry - Meeting Name • 130 - Main Entry - Uniform Title
Added entry	One or more of the following: • 700 - Added Entry - Personal Name • 710 - Added Entry - Corporate Name • 711 - Added Entry - Meeting Name • 730 - Added Entry - Uniform Title
Title Statement	Includes a 245 with a minimum of either $a or $k • 245
Other titles	One or more of the following: • 240 - Uniform Title • 246 - Varying Form of Title • 247 - Former Title
Edition	Includes the following: • 250 - Edition Statement
Version information	One or more of the following: • 254 - Musical Presentation Statement • 255 - Cartographic Mathematical Data • 257 - Country of Producing Entity • 033 - Date/Time and Place of an Event • 034 - Coded Cartographic Mathematical Data
Publication details	One or both of the following: • 260 - Publication, Distribution, etc. (Imprint)

	• 264 - Production, Publication, Distribution, Manufacture, and Copyright Notice
Physical description	Includes the following: • 300 - Physical Description
Other Physical information	One or more of the following: • 306 - Playing Time • 310 - Current Publication Frequency • 321 - Former Publication Frequency • 340 - Physical Medium • 342 - Geospatial Reference Data • 344 - Sound Characteristics • 346 - Video Characteristics • 347 - Digital File Characteristics • 351 - Organization and Arrangement of Materials • 382 Medium of performance • 383 Numeric Designation of Musical Work • 384 Key
Subject	One or more of the following, must have indicator 0 and 7. • 600 - Subject Added Entry - Personal Name • 610 - Subject Added Entry - Corporate Name • 611 - Subject Added Entry - Meeting Name • 630 - Subject Added Entry - Uniform Title • 647 - Subject Added Entry - Named Event • 648 - Subject Added Entry - Chronological Term • 650 - Subject Added Entry - Topical Term • 651 - Subject Added Entry - Geographic Name
Related items	One or more of the following. Must include $a or $t: • 773 - Host Item Entry • 775 - Other Edition Entry • 776 - Additional Physical Form Entry • 780 - Preceding Entry • 785 - Succeeding Entry
Series	One or more of the following. Must include $a: • 490 - Series Statement • 800 - Series Added Entry - Personal Name • 810 - Series Added Entry - Corporate Name • 811 - Series Added Entry - Meeting Name

	• 830 - Series Added Entry - Uniform Title

Application

Quality Level 1 - Breadth:

The Level 1 measure of quality is checking for the presence of the fields shown in each category in the table. When a record has one or more of the fields that are included in a category, that category gets a score of 1. The category scores are then totalled to give the overall quality score.

There are also a couple of general scores:

- A record with no 008 would be deemed to be sub-standard and will only ever get a level 1 score of 1 overall.
- A 'dropped' record will always have a quality score of 0

Quality Level 2 – Depth

The Level 2 measure of quality is checking for the number of the fields, shown in the table above, in two ways:

- A count of repeating fields eg. 650
- A count of multiple fields within a category eg. presence of both a 382 and a 383 in a record

For example, the following subject terms:

650 0$aMissions$zIndochina$xEarly works to 1800
650 0$aMissions$zThailand$xEarly works to 1800
650 0$aMissionaries$zIndochina$xEarly works to 1800
650 0$aMissionariess$zThailand$xEarly works to 1800
650 0$aItalian imprints$xEarly works to 1800
651 0$aIndochina$xChurch history$xSources$xEarly works to 1800

Would generate scores of:

- Level 1: Breadth = 1
- Level 2: Depth = 6

The maximum level 1 score is equal to the number of categories, whilst there is no maximum score for level 2, that is entirely down to the extent of the record.

A set of sample records, with quality scores, is available separately.

Exclusions

The table obviously does not include every possible MARC field, so the reasons for some of the exclusions are given below.

- We are not assessing local fields, so for the purposes of the quality score we will ignore any field with a 9 (eg. 509) other than 490. In the analytics service we can report on the presence of these local fields, if that seems useful, and could also report on local errors such as a field where the institution code in the $5 is not that of the contributor that has sent us the record.

- Leader cp 17 encoding level. The use of # blank to indicate a full level record makes this unreliable.
- 336/7/8. These should be present in every RDA record, so to that extent they are not useful for distinguishing between records, and if a library has had an automated conversion of records to RDA there is a question over how useful these may be as an indication of overall record quality
- 362 Dates of Publication and/or Sequential Designation. There is a question over whether this is/has been used to show generic details about a journal, or local holdings of a journal, and therefore how useful this might be as an indication of overall record quality
- We have excluded a few *very* low use fields, where usage is up to a few hundred records, as their impact would be insignificant and error rates may be high
- 5XX note fields. These may be open to variable usage
- Some 7XX fields. They can be relatively low use and in some cases may show a local focus.
- 880 fields. Not all libraries may see these as of particular significance and it would be possible to express an interest in these by combining a quality score with a requirement for an 880 to be included.
- We are not assessing holdings so we have excluded all 8XX fields, including the 856
- We have not specified a particular style of record, RDA vs AACR2. The intention would be to allow a preference to be specified by combining a quality score with a requirement for an RDA record, where available.