Metadata & Discovery Group

The challenges of data ingest, transformation and aggregation at the National Bibliographic Knowledgebase

Jennie-Claire Crate 0000-0002-6594-8017
Product Manager - Library Hub, Jisc

Received: 08 September 2025 | Published: 22 September 2025

ABSTRACT

Large-scale library data aggregations provide additional discovery opportunities for users and benefit collection management, metadata and interlending work in libraries. However, they present significant delivery challenges around the matching and deduplication of records and require continuous improvement and maintenance to keep pace with changes in the way the sector creates and shares its metadata.

This article describes how contributors to Jisc's National Bibliographic Knowledgebase (NBK) supply their data and how the Jisc Library Hub team matches and deduplicates the data for use via Library Hub Discover, Library Hub Cataloguing, and Library Hub Compare services. It will also explore some of the challenges faced with data transfer, metadata formats, non-standard metadata, and future developments.

This article is based on a paper given at CILIP's Rare Books & Special Collections Group conference in September 2025.

KEYWORDS NBK; metadata transformation; metadata deduplication

CONTACT Jennie-Claire Crate jennie-claire.crate@jisc.ac.uk Jisc

Background

Jisc's Library Hub services were launched in the summer of 2019, replacing the Copac and Copac Collections Management tools created by MIMAS and EDINA's SUNCAT serials catalogue. Underpinning the three services (Library Hub Discover[1], Library Hub Cataloguing[2], and Library Hub Compare[3]) is the National Bibliographic Knowledgebase (known as the NBK). The NBK brings together catalogue data from 205 contributors including national libraries, legal deposit libraries, academic, specialist, and museum libraries. Ingesting, storing and transforming the data supplied by our contributors is a challenging process that requires continual monitoring and adaptation in order to keep up with demand on the service.

[1] https://discover.libraryhub.jisc.ac.uk/
[2] https://cataloguing.libraryhub.jisc.ac.uk/
[3] https://compare.libraryhub.jisc.ac.uk/

Data ingest

One of the chief strengths of the NBK is the breadth of its coverage, and making sure that smaller, specialist libraries are able to contribute their data is crucial to achieving this. Smaller libraries generally have fewer staff members and therefore are sometimes restricted in how they can create and contribute their data in terms of both staff time and technical expertise. This means that the Library Hub team needs to be able to accept data via a variety of methods and support contributors in using them.

The majority of contributor data is sent to the NBK directly via SFTP, but increasingly this method presents problems for libraries. Local IT security measures can prevent implementation and mean that we need to seek alternative methods for data transfer. Additionally, a lack of technical expertise or confidence at the contributing institution can sometimes hamper the setup of SFTP, with the Library Hub team only being able to offer help remotely. In terms of managing the flow of data into the NBK, being able to harvest data from local systems using OAI-PMH would be ideal, but again local cyber security measures now mean that this is increasingly not supported in our contributor's institutions. Cyber security concerns also mean that WebDAV is no longer considered to be an optimal way to transfer data onto our servers, so what other options are there?

It is possible to transfer files by less automated means, such as via online file sharing service like Dropbox or as an email attachment. Whilst seeming to be a solution, these file transfer methods present the Library Hub team with workflow difficulties as we need to manually access, download and then reupload the data to the correct area in our file transfer server. These time-consuming options introduce the possibility of user error on both sides of the transfer, and as a result we cannot support these methods.

To offer a new and more user-friendly option to NBK contributors, the use of Amazon Simple Storage (abbreviated to S3) buckets is currently being trialled with a group of contributing libraries that have experienced difficulties with alternative file transfer methods. If successful, this could work alongside SFTP as one of our main data submission options as it allows the automated collection and processing of data files and therefore fits in with our existing workflows.

Matching and deduplication for Library Hub Discover

Once the data has been received, it is processed and deposited across multiple databases and data stores depending on its intended use. The NBK is just one part of a complex data flow that has to take into account the needs of both users of the three services and those of the Library Hub team and Jisc for monitoring and reporting.

Library Hub Discover, whilst being the most heavily used of the services (with over 10 million searches carried out last year) is perhaps the most straightforward use of the data. As with standard library discovery layers, Library Hub Discover surfaces the metadata held in the underlying database and allows search parameters to be set by

the user, with pre-filtering by data facets such as author, title, format and publication date. However, the scale of Library Hub Discover means that the data needs to be deduplicated to create aggregated records where multiple contributing libraries hold copies of the same item.

Discover's deduplication is carried out using scripts which pass the data through a series of challenges to match and merge identical items. New entries into the NBK are filed alphabetically by title and are then tested against multiple items on either side of their place in the alphabetical index to see whether there are matches in title and multiple standard number fields (ISBN, ISSN, ISMN, ESTC and others). If this doesn't result in a match, metadata relating to edition, creator, date of publication and eventually pagination, score type and map scale are compared. If no match is made, the new entry into the database is treated as a unique item and will be displayed in Library Hub Discover as being held in a single NBK contributing library.

Unmatched records are not usually caused by any kind of issue with the match scripts, but more usually by the variation that is found in the metadata submitted by contributing libraries. In order to enable all sizes and types of library to contribute to the NBK, Library Hub accepts data in almost any digital format from MARC21 to spreadsheets and Word tables. The minimum level of description set for a record is also very low to enable maximum inclusion, and only a record ID number and item title are mandated. What this means in practice is that these less-full, lower quality descriptions have far fewer data points against which to match, making it much more likely that they will be assessed to be "unique" even where they are not. In these cases, users of Library Hub Discover will find multiple entries for what appear to be identical publications which can cause frustration when searching for the nearest copy of a book for researchers and uncertainty for interlibrary loan teams looking for an item for their patrons.

Matching and deduplication for Library Hub Compare

Deduplication issues also have an impact on users of Library Hub Compare, Jisc's collection management tool. Library Hub Compare uses the same deduplicated index as Library Hub Discover, allowing library teams to compare their local holdings with other NBK contributor institutions. This collaborative collections management approach is about to come to the fore for libraries in the UK with the launch of the UK Print Book Collection (UK PBC), an initiative created to ensure that a minimum of seven print copies of books published in or before 2010 is retained in the UK whilst enabling libraries to make evidence-based decisions when managing their physical collections. UK PBC will launch in October 2025, and is using Library Hub Compare to show libraries which items in their print collection might need to be retained to avoid losing one of the final seven remaining copies. The benefit of this work to libraries is that they can safely manage down their print holdings and repurpose physical space in their buildings to meet the needs of their users more effectively, and for library users it means that access to the print items they need will be preserved. Just as with Library

Hub Discover though, poor or sparse metadata means fewer matches being made and therefore more library holdings being reported as ‘unique’. This results in lower confidence in the Compare reports for UK PBC and can cause confusion for collections management teams. All collection analytics tools tend to over-report item rarity due to deliberately conservative matching algorithms, which prioritise avoiding false positives over identifying duplicates. Making incorrect matches is even more undesirable than incorrectly labelling an item as rare or unique, but clearly neither is giving a completely accurate picture.

An additional challenge for Library Hub Compare is the currency of the dataset. Contributors choose how often to update their holdings in the NBK. Most opt to do this weekly or monthly but longer intervals can be caused by lack of staff capacity or changes of key team members in contributing libraries, or by the disruption caused by a change of library management system. Data currency is important for collections management analytics as it ensures decisions are being made based on the most recent available data, bringing confidence to stock management and editing activity. This is another strong argument for preferring automated data transfer methods, as once they have been set up and scheduled they can be carried out frequently with minimum staff intervention, ensuring data currency and workflow efficiency.

Ways forward

Libraries have historically tailored metadata practices to meet local needs, system requirements, and workflows. In addition to this, the hybrid environment created by changes in standards that are not universally adopted means that it can be difficult to define what “good” metadata looks like in the NBK and adds another layer of complexity to deduplication in the database. Revisiting and reassessing local metadata practices that vary from the standard and improving minimal legacy records could go a long way to addressing issues with data matching in large aggregations but of course comes with a cost to libraries. The benefits of this work, and the retrospective application of the new standards, can be difficult to articulate in a business case but could bring financial benefits in the long term when it becomes possible to download and ingest data from the NBK without having to edit it to match local practice. Moving toward more standards-based approaches can facilitate collaborative description and collections management, simplify system migrations, and ease the adoption of new tools and evolving standards. While innovations like linked data can be slow to take root, the long-term benefits in enhancing discoverability via the semantic web and for accessibility are substantial for both collections teams and users.