

# Maximizing Efficiency in Shared Print Programs: A DIY Approach to Overlap Analysis

Marlene van Ballegooie

## Abstract

This paper examines the use of the KNIME Analytics Platform as a do-it-yourself solution for collection overlap analysis in shared print programs. Focusing on the Keep@Downsview partnership and the Canadian University Press Project, we demonstrate how KNIME significantly enhances the efficiency and accuracy of title matching and metadata processing in library operations. By outlining the methodologies used in these workflows, we highlight KNIME's effectiveness in analyzing title overlaps in diverse library collections. This study reveals the transformative potential of KNIME in collaborative collection management, offering valuable insights into its effectiveness and practicality for streamlined overlap analysis.

**Keywords:** shared print programs, collaborative collections, overlap analysis, KNIME Analytics Platform

## Introduction

As academic libraries face challenges such as physical space constraints, limited funding, and evolving resource usage, collaborative

collections and shared print programs have gained prominence. These initiatives not only promote inter-institutional cooperation but also optimize resource allocation and ensure the long-term preservation and accessibility of scholarly materials. By participating in shared print programs, libraries can significantly broaden access to diverse resources, thereby more effectively supporting the research and educational needs of their academic communities. A critical challenge in implementing shared print programs is conducting efficient and accurate overlap analysis to identify duplicate titles across diverse collections. To address this challenge, the University of Toronto Libraries (UTL) initiated a project leveraging the KNIME Analytics Platform—an open-source, low-code software solution. This paper presents the methodology developed to streamline overlap analysis, focusing on efficient title matching and metadata processing techniques that enhance accuracy while reducing the workload for library staff. Centered on the Keep@Downsview partnership and the Canadian University Press Project, this research demonstrates KNIME's potential to transform data analysis in collaborative collections, offering valuable insights into its capabilities and practical applications for efficient overlap analysis.

## **The Keep@Downsview Shared Print Collection**

The Keep@Downsview shared print program highlights the importance of collection analysis in shared print initiatives. This collaborative effort involves the University of Toronto, the University of Ottawa, Western University, McMaster University, Queen's University, and Memorial University. The primary goal of this partnership is to preserve the scholarly record by utilizing a shared high-density storage and preservation facility at the University of Toronto's Downsview Campus in North Toronto. Overlap analysis plays a vital role in optimizing resources and ensuring effective preservation

within this program. To maximize storage efficiency and uphold preservation goals, program partners work to prevent duplicate physical resources from being included in the facility. When a partner library identifies that a copy they intend to preserve is already housed at the Downsview facility, they can withdraw their own copy and share ownership of the preserved copy. This strategy enables partners to focus their efforts on preserving unique materials. Central to the shared print partnership's mission is the principle of sharing, not duplicating, resources within the preservation facility. As such, the efficient identification and elimination of duplicates from the collective collection is essential for the success of the Keep@Downsview program.

## **Challenges in Overlap Analysis**

Despite the critical importance of overlap analysis in shared print initiatives, libraries face limited options for effective and affordable tools. Current tools range from well-known solutions like OCLC's GreenGlass and the Colorado Alliance of Research Libraries' Gold Rush Library Content Comparison System, to more basic alternatives such as Microsoft Excel. Another challenge for overlap analysis is that collections are dynamic, requiring regular analysis to accurately capture the current state of the collection. Further complicating the process is the variability in metadata across institutions. Identical resources may be assigned different identifiers, such as ISBNs or OCLC numbers, making duplicate title detection a challenging task. The evolving landscape of bibliographic standards and cataloging practices adds another layer of complexity to the analysis. These challenges underscore the pressing need for innovative and adaptable solutions that can streamline overlap analysis and enhance the efficiency of collaborative collections.

## **The KNIME Analytics Platform**

To ease the burden of overlap analysis, staff at UTL began to explore software beyond traditional library tools. In this investigation, the KNIME Analytics Platform emerged as a highly promising solution, offering essential functionality for conducting large-scale data comparisons. KNIME provides easy access to data, enabling users to combine, analyze, and visualize information without requiring coding skills. Its low-code, no-code interface makes analytics accessible to users of all skill levels. A significant advantage of KNIME is that it is free, requiring only an investment of time to learn and effectively utilize its capabilities.

The KNIME Analytics Platform offers an extensive range of pre-built nodes, each designed to perform specific actions on data. These actions include tasks such as data reading, joining, partitioning, visualization, and model training, among others. Users can easily construct workflows by connecting nodes through an intuitive, drag-and-drop graphical interface. One significant advantage of using the sequential workflow approach in KNIME is the clarity it provides at each step of the analysis process, minimizing the time required to identify and rectify errors. Given its adaptability and functionality, KNIME holds great potential for supporting data-driven processes within libraries.

## **Use of KNIME in the Keep@Downsview Shared Print Program**

In developing a new approach to collection analysis, the strategy involved leveraging existing systems and tools while expanding functionality as needed. Within the Keep@Downsview partnership, each participating institution uses the Alma library services platform for inventory management, which includes Alma Analytics for reporting.

Bibliographic and operational data from Alma undergo a daily Extract, Transform, Load (ETL) process before being loaded into Alma Analytics, making it an ideal starting point for overlap analysis. To ensure compatibility across partner libraries, we created a standardized report including selected data fields for extraction. Partners simply need to scope the report by specifying the system identifiers for the collection overlap. Separate files containing relevant ISBNs or ISSNs are also provided with the bibliographic data. Once extracted, the data are shared with the University of Toronto for integration into the KNIME workflow.

Within the KNIME overlap workflow, library data is managed in two streams: one containing title-level data for the entire UTL collection, and the other containing partner title data for comparison. To initiate the process of identifying overlaps, the data required further transformation to generate the necessary match keys. Previous attempts at overlap analyses relied on matching standard bibliographic identifiers like OCLC numbers, ISBNs, or ISSNs. Although this approach is valid for determining matches at the title level, it fails to account for situations where similar titles may possess different identifiers or where bibliographic records lack them entirely. Therefore, it was essential to develop additional match keys that could accurately identify similar titles across collections.

The concept of generating match keys from bibliographic data originated with the Colorado Alliance of Research Libraries, specifically through their development of a matching algorithm for the Gold Rush Library Content Comparison System. Unlike conventional approaches that rely solely on standard identifiers, this system derives a single match key from a combination of elements within the bibliographic record.<sup>1</sup> The principle behind this approach is that libraries universally adhere to bibliographic standards, ensuring consistency in their records. This standardization means that match keys derived from bibliographic data should be identical across different library systems when referring to the same item. In our KNIME experimentation, we expanded on this idea by generating multiple match keys with varying levels of specificity, rather than a single match key. This

strategy significantly increased the probability of identifying matches between the partner data and the University of Toronto Libraries collection.

To achieve precise matches in the overlap analysis, we implemented a comprehensive approach to maximize the potential for accurate identification. Using various KNIME nodes for string manipulation, we normalized the data elements and combined them in different configurations to create match keys, allowing for flexible and robust overlap analysis. This standardization process involves removing capitalization, punctuation, spaces, diacritics, and special characters. Additionally, to increase the likelihood of identifying common materials across collections, fragments of the bibliographic elements are used instead of the entire data element. The match keys generated through this normalization process are illustrated in Table 1.

After generating match keys for both the University of Toronto Libraries collection and the partner data, the matching process begins. The KNIME workflow uses “joiner” nodes to identify matches between the collections. To ensure precision and streamline the results, a tiered approach to matching is implemented. As shown in Table 1,

**Table 1. Match Key Examples**

| Match Key Type                   | Match Key Rank | Match Key Example                                                                            |
|----------------------------------|----------------|----------------------------------------------------------------------------------------------|
| OCLC Number                      | 1              | 36720114                                                                                     |
| ISBN                             | 2              | 9780306464072                                                                                |
| ISSN                             | 3              | 08466629                                                                                     |
| LCCN                             | 4              | 81011585                                                                                     |
| Title/Author/Date/Publisher      | 5              | touchingbaseprofessionalbaseballand<br>americancultureintheprogressive<br>era-ries-1999-univ |
| Title/Author/Date                | 6              | modernmethodsforcomputersecuritya<br>ndprivacy-hoff-1977                                     |
| Title/Author                     | 7              | governingaftercommunisinstitutions<br>andpolicymaking-dimi                                   |
| Truncated Title/Author (6 words) | 8              | heideggerandmarxaproductivedialo<br>gue-hemm                                                 |
| Truncated Title/Author (5 words) | 9              | beyondhumanismessaysinthe-hart                                                               |
| Normalized Title                 | 10             | explorationsinsociologyandcounseling                                                         |

each match key is ranked by confidence level. The OCLC number is the highest confidence match, as it uniquely identifies records in the WorldCat database, the largest network of library holdings worldwide. Other reliable matches include bibliographic identifiers like ISBNs, ISSNs, and LCCNs. Lower confidence match keys are based on single data elements, such as a normalized title or a truncated version of it. In the KNIME workflow, match keys are processed sequentially, and once a match is found, the corresponding title is excluded from further analysis. This systematic approach ensures a thorough and accurate identification of matches.

Once matches are identified, the titles undergo a secondary analysis to determine the appropriate actions for the partner library regarding the physical resource. This analysis checks whether the title is held at the Downsview preservation facility. If the Downsview facility holds a copy of the title, the "SHIP\_OR\_SHARE" field in the results is set to "SHARE," indicating that the partner library can deaccession the title from their collection and share ownership of the existing copy held at Downsview. Conversely, if Downsview is not listed among the holding libraries, the "SHIP\_OR\_SHARE" field is set to "SHIP," indicating that the partner library needs to ship the physical volume along with its metadata to Downsview. The "SHIP" or "SHARE" designation provides significant time savings for partner libraries, promptly informing them of the required actions to facilitate processing by the University of Toronto Libraries.

In addition to indicating whether the matching title will be shipped to Downsview or shared, the results also include a MATCH\_TYPE field that specifies the match key used during the title overlap identification process. This additional data point allows partner libraries to quickly assess the confidence level associated with each match. After evaluating all match keys, the resulting lists of title overlaps are merged into a comprehensive list and exported to an Excel spreadsheet. Titles that did not match any resource in the University of Toronto catalog are saved in a separate tab within the same spreadsheet for easy reference.

## **Impact of the KNIME Overlap Tool**

As a key component of the planned deaccessioning initiatives by Keep@Downsview partners, the KNIME Analytics Platform has significantly enhanced the efficiency and accuracy of overlap analysis within the shared print program. By automating the title matching process, KNIME has dramatically reduced the staff time required to compare collections against the University of Toronto Libraries holdings. For partner libraries, the platform not only identifies overlaps but also details the criteria for each match, instilling greater confidence in the results. High-confidence matches, such as those based on OCLC numbers, can be approved without manual review, while lower-confidence matches, like those derived from normalized titles, prompt further examination. Moreover, the tool specifies which UTL library holds the corresponding copy, simplifying decisions on whether items should be transferred to the Downsview facility. If a partner library identifies a copy they wish to preserve already stored at Downsview, they can withdraw their own copy and share ownership of the preserved one. By eliminating much of the guesswork from metadata processing, the overlap analysis tool significantly optimizes partner workflows and streamlines the preparation of materials for long-term preservation at Downsview.

## **Application of KNIME in the Canadian University Press Project**

To further showcase the capabilities of the KNIME Analytics Platform for collection analysis, we made key enhancements to the overlap workflow in the Canadian University Press (CUP) Project initiated by NORTH/NORD: The Canadian Shared Print Network. NORTH/NORD is a collaborative Canadian initiative that coordinates the activities of existing regional shared print initiatives to support the preservation of print collections and ensure long-term accessibility for users. The

CUP Project aims to preserve monographs published by seventeen Canadian University Presses from across Canada. The goal of this project is to identify widely and scarcely held titles across academic and government libraries and secure retention commitments for three copies of each item—one for preservation and two for access. Three key features set this KNIME overlap workflow apart from the Keep@Downsview workflow: the ability to compare multiple title lists simultaneously, the integration of AI-generated Python scripts for ISBN clustering and fuzzy title matching, and the automatic assignment of retention commitments to partner libraries.

The CUP workflow builds on the foundation of the Keep@Downsview overlap tool, enhancing and adapting its functionality to address the distinct requirements of the project. This workflow is specifically designed to compare multiple title lists against each other to determine a core list of unique titles for each university press. The workflow was developed in response to the challenge that many university presses were unable to provide comprehensive lists of all titles they had ever published. Consequently, the KNIME overlap workflow was adapted to create a core list based on data supplied by various libraries. For each university press, a separate KNIME workflow is created, where the title data is consolidated into a single table for processing. As titles progress through the workflow, various match keys are used to compare entries, merging duplicates until only a unique list of titles remains. This approach ensures a comprehensive and accurate compilation of unique titles for each university press.

To further enhance the CUP workflow, AI tools were used to develop Python scripts that serve two critical functions: clustering ISBN families and conducting fuzzy title searching. Using ChatGPT, a script was developed to cluster ISBNs, allowing titles to be considered matches even if they did not directly share an ISBN but were connected through a network of shared ISBNs. For example, if Title A shares an ISBN with Title B, and Title A shares another ISBN with Title C, the script clusters Titles A, B, and C together. Additionally, we employed AI for fuzzy title matching. ChatGPT facilitated the creation of a robust matching

algorithm that evaluates word and character-level similarities, normalizes publication dates, and compares page numbers. By implementing fuzzy matching techniques, the algorithm can successfully identify matching records even when faced with minor discrepancies in titles, author names, or other bibliographic details. The seamless integration of these AI-driven approaches significantly enhanced the accuracy and efficiency of title matching, particularly in cases where records contained minor variations, typographical errors, or other inconsistencies. This improvement not only increased the overall match rate but also reduced the need for manual intervention, streamlining the overlap analysis process for the CUP project.

Another key enhancement of the CUP workflow is the automatic assignment of retentions to partner libraries. KNIME automates this process by integrating logic to ensure each unique title secures three retentions. For each university press, the holding libraries are ranked to ensure that retentions are appropriately distributed and optimized. As each title progresses through the workflow, the first retention is allocated to the home university press, the second aims to secure a preservation copy at Library and Archives Canada, and the third is assigned to another library based on criteria such as the presence of a preservation facility, geographic location, and availability. This automated, rule-based approach significantly streamlines the decision-making process for book retention, enhancing efficiency and ensuring comprehensive coverage for preservation.

## **Challenges and Future Directions**

Although the KNIME Analytics Platform has proven effective in the overlap analysis of monographs, certain formats—particularly serials—present significant challenges for title matching. Serials are inherently complex due to their dynamic and ongoing nature. Often among the oldest records in library collections, serials require consistent maintenance over time, and evolving metadata standards add

further complexity. Current cataloging practices use the successive entry convention, creating new bibliographic records for each change in title or corporate body, while older records may follow the previous latest entry convention, where serials are cataloged under the most recent title with former titles noted in free-text fields. These shifts in cataloging rules, coupled with varying interpretations and local practices across institutions, result in significant metadata inconsistencies. The problem is further compounded by generic titles such as *Annual Report*, *Proceedings*, or *Bulletin*, which make accurate identification and matching particularly challenging. Moreover, the inconsistent application of standard identifiers like OCLC numbers and ISSNs exacerbates these issues, making reliable overlap analysis for serials difficult to achieve with KNIME.

Music recordings, scores, and electronic resources also present unique obstacles for overlap analysis. The non-unique nature of titles for music recordings and scores often leads to difficulties in distinguishing between different items. Additionally, multiple versions of the same work complicate the matching process, as the match keys used in the analysis are often not distinct enough to confirm a match with certainty. Similarly, electronic resources pose significant challenges due to the poor quality of metadata in e-resource knowledgebases, which hinders accurate matching. The bibliographic records for electronic resources frequently do not describe the exact version of the resource, leading to potential mismatches. This lack of precise metadata makes it difficult to identify and verify electronic resources in overlap analysis, highlighting the limitations of using KNIME for these specific formats.

The future of the KNIME overlap analysis tool is focused on continuous refinement and expansion to enhance its capabilities and address emerging challenges. For the Keep@Downsview partnership, we will direct efforts towards fine-tuning the workflow and match keys as more title lists are processed, aiming for increasingly optimal results through iterative improvements. The North/Nord Canadian University Press project has reached a significant milestone, with over 270,000

titles analyzed and retention libraries identified. The next steps involve reviewing the results for each university press, correcting any discrepancies, and distributing the updated title lists to participating libraries for the application of retention notes in their catalogs. Once the initial work with the Canadian University Press project is complete, we will explore the feasibility of a Phase 2, which will analyze print title lists against available electronic versions, further broadening the scope and impact of the KNIME overlap tool.

To enhance the preservation workflows at the University of Toronto Libraries, we aim to adapt the KNIME workflows to effectively identify item level overlaps among resources within our library system. This adaptation will streamline the manual verification process necessary before transferring materials to the Downsview preservation facility. Additionally, we plan to expand the KNIME workflows' capabilities to detect overlaps between physical and electronic records within our library system. This enhancement will provide a more comprehensive view of our collection, enabling more informed decision-making in resource management. By implementing automated methods to identify duplication, we will significantly reduce the time-consuming manual labor involved in our collection management processes.

## **Conclusion**

The implementation of the KNIME Analytics Platform for overlap analysis has proven transformative for both the Keep@Downsview partnership and the Canadian University Press project. By automating critical aspects of the overlap analysis process, KNIME has significantly enhanced the efficiency and accuracy of metadata work for these shared print initiatives. As we continue to refine and expand the capabilities of our KNIME workflows, our focus remains on addressing emerging challenges and optimizing the tool's performance. Future developments will prioritize enhancing match key precision, incorporating the analysis of electronic resources, and exploring additional

applications within the University of Toronto Libraries system. These advancements will support our goals of preserving scholarly materials, enhancing collaborative collections, and serving the broader academic community more effectively. As we continue to innovate and adapt, KNIME will remain a cornerstone in our efforts to advance library operations, promote sustainable practices in collection management, and ensure long-term preservation and access to scholarly materials.

## Contributor Notes

**Marlene van Ballegooie** is the Metadata Technologies Manager, University of Toronto Libraries, Toronto, Ontario, Canada.

## Note

- 1 George Machovec, "Shared Print Analysis Tool at the Colorado Alliance of Research Libraries," *Collaborative Librarianship* 8, no. 1 (2016): 29–40, <https://digitalcommons.du.edu/collaborativelibrarianship/vol8/iss1/7>.