11-14-2023

Migrating 120,000 Legacy Publications from Several Systems into a Current Research Information System Using Advanced Data Wrangling Techniques

Yrjö Lappalainen
Zayed University

Matti Lassila
Tampere University

Tanja Heikkilä
Finnish Geospatial Research Institute

Jani Nieminen
Tampere University

Tapani Lehtilä
Tampere University

publications

MDPI

Article

Migrating 120,000 Legacy Publications from Several Systems into a Current Research Information System Using Advanced Data Wrangling Techniques

Yrjö Lappalainen [1,*], Matti Lassila [2], Tanja Heikkilä [3], Jani Nieminen [2] and Tapani Lehtilä [2]

[1] Library and Learning Commons, Zayed University, Dubai P.O. Box 19282, United Arab Emirates
[2] Tampere University Library, Tampere University, 33014 Tampere, Finland; matti.lassila@tuni.fi (M.L.); jani.nieminen@tuni.fi (J.N.); tapani.lehtila@tuni.fi (T.L.)
[3] Finnish Geospatial Research Institute (FGI), National Land Survey of Finland (NLS), 02150 Espoo, Finland
* Correspondence: yrjo.lappalainen@zu.ac.ae

Abstract: This article describes a complex CRIS (current research information system) implementation project involving the migration of around 120,000 legacy publication records from three different systems. The project, undertaken by Tampere University, encountered several challenges in data diversity, data quality, and resource allocation. To handle the extensive and heterogenous dataset, innovative approaches such as machine learning techniques and various data wrangling tools were used to process data, correct errors, and merge information from different sources. Despite significant delays and unforeseen obstacles, the project was ultimately successful in achieving its goals. The project served as a valuable learning experience, highlighting the importance of data quality and standardized practices, and the need for dedicated resources in handling complex data migration projects in research organizations. This study stands out for its comprehensive documentation of the data wrangling and migration process, which has been less explored in the context of CRIS literature.

Keywords: current research information system (CRIS); research information; data migration; legacy data; data quality; machine learning; data wrangling; natural language processing (NLP)

Citation: Lappalainen, Y.; Lassila, M.; Heikkilä, T.; Nieminen, J.; Lehtilä, T. Migrating 120,000 Legacy Publications from Several Systems into a Current Research Information System Using Advanced Data Wrangling Techniques. *Publications* **2023**, *11*, 49. https://doi.org/10.3390/publications11040049

Academic Editor: Costantino Thanos

Received: 28 August 2023
Revised: 27 October 2023
Accepted: 9 November 2023
Published: 14 November 2023

1. Introduction

In 2019, two universities in Tampere merged to create Tampere University, the second-largest university in Finland. The new foundation-based university also became the main owner of Tampere University of Applied Sciences, a professional higher education institution oriented towards working life and RDI co-operation. With more than 30,000 students and 5000 employees, these organizations now form the higher education community of Tampere Universities (TUNI) with a competitive edge in technology, health, and society. Due to the merger, the community had to re-tender all of its IT systems, which led to a large number of system procurement and implementation projects within the community.

Before the merger, a project was launched to procure a single current research information system (CRIS), aiming to harmonize the registration, validation, reporting, and presentation of research outputs and other research information within the university community. In addition to the university community, a local hospital district (Pirkanmaa Hospital District) also joined the project. The Tampere University Library had already provided library services and validated publications for the hospital district for a number of years. Although there was significant overlap between the publications of the university and the hospital district (many faculty members are affiliated to both organizations), the hospital district used a separate system to register their publications.

In the beginning of the project, the university community and the hospital district had three different systems to maintain their publication data. Since the organizations also had different guidelines for data collection and the data had been collected over a long period

of time, the quality of metadata varied greatly. This was not considered a major issue in the beginning of the project since the system provider was supposed to handle all data migrations. However, the provider unexpectedly declined to perform any data migrations, and the daunting task was left entirely to the project team. Since the plan was to also migrate all historical data (including decades of publication history), the total number of publication records was around 120,000.

As the project progressed, it became clear that the legacy publication data were highly diverse and poorly structured. In particular, the author data became a major issue since they contained various delimiters between names, and the format had changed over the years. Handling this kind of data manually would have been an impossible task in terms of workload, and therefore the project team decided to use advanced data wrangling methods (notably, the probablepeople Python library) to automatically process the data. However, a great deal of manual work was still needed due to the unstructured nature of the source data. In total, about 335,000 name occurrences were processed, of which about 10,000 were manually corrected. In addition, another issue was to merge a large number of duplicate records originating from different systems, due to a significant overlap between the former systems. This article describes the eventual success of processing and merging a large number of publication data from various sources into a single CRIS. While focusing on data processing and quality, this article also discusses this complex CRIS implementation in general, and serves as a constructive example for similar migration projects that involve large quantities of unstructured and duplicate data.

2. Literature Review

In Finland, publications play a crucial role in national research funding. In the current funding model (valid from 2021 to 2024), 14% of the funding received from the Finnish Ministry of Education and Culture is based on publications. To assess the quality and impact of publications, national panels of experts classify all publication channels, including journals, publication series, and publishers, using the Publication Forum classification system. The funding allocation is influenced by the quality, format, and target audience of the publication, with scientific peer-reviewed articles in high-quality journals attracting the highest funding. Open access publications are also rewarded with a coefficient of 1.2 [1,2]. All publication data are reported to a national publication service called Virta, and are in turn publicly displayed in the Research.fi portal [3]. As a result, the metadata must comply with rigorous national requirements for publication data collection.

While most universities regularly harvest their publications from international databases such as Scopus and Web of Science, the Ministry of Education and Culture also requires additional metadata that are not available in any database. Furthermore, many Finnish language publications and conference papers are not indexed in these international databases, requiring researchers to manually input them into local systems first. Because of this, all metadata must be validated and supplemented before they are reported to the Ministry of Education and Culture. Generally, university libraries in their respective organizations are responsible for validating and reporting all publication data.

The trend towards comprehensive CRIS platforms began in the early 2000s, although many universities and research organizations had already used various research registers for years prior to this. In Finland, the first computerized research register was established in the 1970s, and a government committee was set up to plan a national research register as early as 1974. Although the national register did not materialize at the time, individual research organizations established their own registers, with 40 systems in operation already in 1987 [4]. In 1989, Tampere University of Technology implemented its first research register which served faithfully until it was replaced by Pure CRIS in 2015. The University of Tampere had also had its own research register since the late 1990s.

The CERIF (Common European Research Information Format) model was released in 1991 with a goal to establish a common data format for the exchange of research information. EuroCRIS, an organization promoting CRIS best practices, had existed informally since the

first international CRIS conference in 1991. In 2004, EuroCRIS was formally re-founded and registered as a not-for-profit organization in the Netherlands [5]. Around the same time, the development of Pure and Converis began, marking the emergence of prominent CRIS platforms.

Early research articles and reports on current research information systems provided introductions to the concept of CRIS platforms and the CERIF data model. Joint [6] discussed the role of CRIS platforms in managing research information, and how libraries could use them to play a more effective role in the research process. The article highlighted the success of the CRIS concept in the Netherlands, where the National Academic Research and Collaborations Information System (NARCIS) portal incorporated the entire network of digital academic repositories in the country.

In a systematic literature review of CRIS articles published between 2007 and 2017, Velásquez-Durán and Ramírez-Montoya [7] identified four major topics during this period: (1) data structure, which outlined the nature of CRIS, its data types, and data structure, (2) applications, which discussed the ways in which CRIS-managed information is utilized; (3) communication/interoperability, which addressed topics such as standards, protocols, and identifiers that are intended to facilitate the integration of CRIS with other systems; and (4) institutional repository, which focused on the relationship between CRIS and an institutional repository, and the challenges of merging the two systems.

Interoperability and standards such as the CERIF data model [8], OpenAIRE compatibility [9–12], and the relationship between CRIS platforms and institutional repositories [13] have been major topics of interest in CRIS literature. Since 2017, research measurement and research data management [14,15] have also gained more attention in the context of CRIS platforms. More recently, there has been a growing trend of establishing national CRIS platforms and research portals, which have emerged from various countries, including the Netherlands [16], Norway [17], Denmark [18], Finland [3], India [19], Portugal [20] Ukraine [21], Brazil [22], Croatia [23], Czech Republic [24], and Türkiye [25]. Some of these systems operate as shared standalone instances, while others function as centralized platforms that aggregate information from individual systems on a national level. In 2022, EuroCRIS also launched a new working group called CRISCROS, with the aim of bringing together representatives from various national and regional CRIS initiatives in Europe and beyond [26].

The importance of data quality, particularly when it comes to the integration of data from other systems into CRIS, has received much attention in CRIS literature (see, e.g., [27–30]). CRIS platforms generally collect information from various internal and external sources, such as HR systems, financial systems, student registries, facility and equipment registries, etc. These sources and their data structures are often heterogeneous, maintained by multiple stakeholders, and designed for various purposes. In order to integrate and use these data effectively, the data need to be cleaned, transformed, harmonized, and merged before loading them into CRIS. This generally involves setting up a separate integration platform or a "staging area" to complete the transformations [29].

Although the quality of research information in general has been widely discussed, and data quality has been a major issue in many CRIS implementation projects (see, e.g., [30–34]), issues related to the migration of legacy publication data have remained surprisingly overlooked. Bevan and Harrington [35] described lessons learned from implementing a CRIS platform at Cranfield University, which was accomplished within a year using Converis. Data quality was an issue, as the team combined publication records from three separate systems with duplicate items. There was overlap between the systems, and the structure and quality of data also varied. The authors reported that ideally the team should have spent more time checking and cleaning the data prior to importing them in order to optimize the consistency of the data, but did not elaborate on how the data quality issues were resolved.

Buchmayer et al. [31] reported a Pure implementation at the University of Vienna. Although not focusing on data quality, the authors noted that the implementation of master data synchronization involved a higher-than-expected workload, and the migration of

existing data led to problems that had not been considered before. Siciliano et al. [32] reported a Converis implementation at the Free University of Bozen-Bolzano, in which the project team migrated publication data from a previous database with a history of 17 years "in order to achieve a faster and easier acceptance of the new system". The authors noted that this significantly increased the workload for the implementation team, but did not elaborate on the steps taken to process the legacy data.

Grenz et al. [33] described a CRIS implementation at King Abdullah University of Science and Technology (KAUST) in Saudi Arabia. According to the authors, a great deal of intermediary development was required before information from various source systems could be structured and ingested into Pure in a usable form. Furthermore, a custom solution was necessary to make the integration between Pure and DSpace. As a new university, KAUST did not have an extensive publication history of its own. However, the project team wanted to include individual pre-KAUST publication histories into Pure. Since Pure uses Organizations and Persons as the basis of all other data, importing external publications became an issue because the prior affiliations were being replaced with their current affiliation at KAUST. This raised data quality issues and delayed the entire implementation project. After facing various challenges with Pure, KAUST decided to adopt a more system-neutral approach where the reporting, preservation, and analysis of research information were distributed across different systems (Pure, DSpace, PlumX, and a locally developed ORCID integration).

Van den Berghe and van Gaeveren [30] reported a migration of approximately 90,000 publication records from a legacy system into Pure at Vrije Universiteit Brussel. Choosing a different approach, they decided to opt for "a minimal cleaning pre-migration and setting up a large-scale cleaning effort post-migration". They prioritized data quality efforts by giving a lower priority to tasks that are complex and have little use compared to simpler tasks that result in cleaner and more reusable data. They also performed reactive data quality assessment and improvement based on feedback from researchers and administrators. According to the authors, analyzing at the field level rather than the record level allowed for better resource allocation and planning. Duplicate detection was facilitated by the tools provided by Pure, allowing for the quick detection and merging of duplicates. However, due to the absence of sufficient bulk editing tools in Pure, the authors developed an automated data-cleaning solution using Python and the Selenium browser automation tool to facilitate data entry and record manipulation. The authors highlighted an example where a specific field in 15,000 records needed to be changed. They estimated that one person could have manually processed about 300 records per day, but the automated script was able to process the entire batch within 24 h.

3. Procurement Project

Before the project, the Tampere university community and the Pirkanmaa Hospital District used three separate systems to register their research information: TUTCRIS (Elsevier Pure) at the former Tampere University of Technology, SoleCRIS at the former University of Tampere, and Julki at the Pirkanmaa Hospital District. SoleCRIS and Julki are provided by Finnish vendors. Each system had its own administrators and content validators, as well as varying guidelines for data collection.

In 2018, a project was launched to procure a single CRIS platform for the Tampere university community and the hospital district, aiming to harmonize the registration, validation, reporting, and presentation of research output and other research information. The goal was also to enhance the visibility of research activities, promote open access, and contribute to the internal and external customer relationships and partnerships of the community by presenting its experts, units, and infrastructure. The combined number of research and teaching staff in Tampere University and the Pirkanmaa Hospital District is about 2300 persons, and the estimated number of research outputs is about 6000 publications per year.

The invitation to tender was launched in December 2018. The scope included the implementation of the system, data migrations, configurations, integrations, testing, maintenance, and training services. The system was requested to be provided as a hosted solution that would be maintained by the successful tenderer. The goal was to procure a modern system that had comprehensive reporting functionalities and enabled a high level of integrations with local systems and external databases. Modern CRIS platforms support manual data collection by integrations with international databases (e.g., Scopus and Web of Science) and global researcher IDs (e.g., ORCID), reducing the need for manual input. They can also be integrated with institutional repositories and other external services such as SciVal and Sherpa/ROMeO, and altmetric tools such as PlumX.

During the university merger, the Tampere university community established a new central integration platform that collects data from all key systems at the university community, such as the HR system, student registry, identity management system, organization registry, and project/application registry. The university community owns and manages the platform and all system integrations are carried out through it, which makes it easier to move data between key systems. The integration platform and planned CRIS integrations and data flows at the start of the project are depicted in Figure 1.

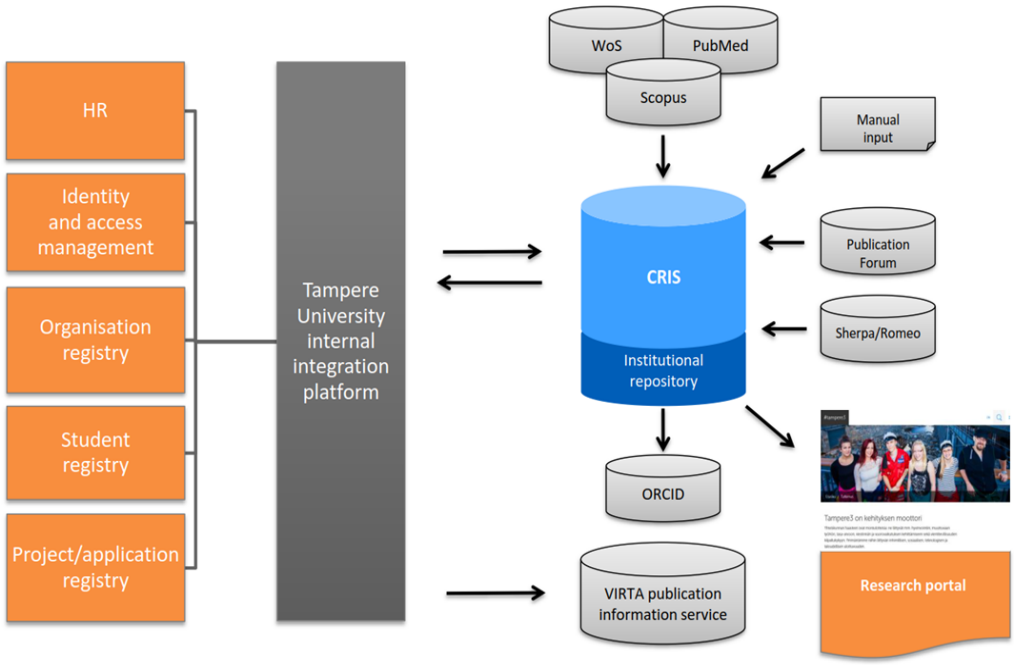

Figure 1. Overview of planned CRIS integrations and data flows at Tampere University at the start of the project.

To ensure a high level of integration and functionality, the system requirements consisted of more than 300 technical and functional requirements. A usability test was also part of the award criteria. After a long period of negotiations, the contract was awarded to Elsevier (Pure). However, the tendering process had been significantly delayed from the original timeline, which also resulted in time constraints and urgency during the implementation project.

4. Implementation Project

Once the procurement was complete, a follow-up project was launched for the CRIS implementation in 2019. The project had the following main goals:

1. Implement a new common system and related integrations;
2. Migrate all existing data from the current systems into the new one and shut down the old systems;

3. Provide the university community with excellent capabilities for collecting, validating, reporting, and presenting research information.

The original plan was divided into three phases: preparation, implementation, and deployment. The plan had an ambitious schedule to begin with, as the implementation was supposed to be fully completed by summer 2020 (including the decommissioning of the old systems). The original plan consisted of the following steps:

- Designing integrations and data model—October 2019;
- Mapping existing background information, publication data, and other research data—October 2019;
- Deployment of the test server—November 2019.
- Implementation of integrations—January 2020;
- Exporting background information (researcher profiles, units, identifiers) to the new system—January 2020;
- Exporting existing publication data and other research data to the new system—February 2020;
- Installation and configuration of the public research portal—March 2020;
- Testing the system and verifying the data—May 2020.
- Deployment of the production server—May 2020;
- Decommissioning of the old systems—August 2020.

In the project's risk analysis, concerns were raised about system integrations (e.g., HR system, identity management, student registry, etc.) and the availability of the university's integration team. However, the quality of legacy publication data or the workload caused by the migration was not identified as a risk beforehand at all, because we were under the impression that the system provider would handle the migration.

After the procurement contract had already been signed and the implementation project was underway, it came as a surprise to the project team that we would have to perform the entire data migration on our own. The heterogeneous quality of the source data had not yet been identified as a risk. As a result, the lack of technical capabilities (i.e., readiness for data processing) was not identified as a risk either. If these risks had been identified, the project team could have had the opportunity to learn the use of new tools during the planning phase and to distribute the workload of data cleaning and preparation among multiple people. In retrospect, the quality of source data should have been thoroughly evaluated before starting the project, and the project schedule should have been re-planned before committing to the data migration task. On the other hand, there was an urgent need to move away from overlapping systems to a new CRIS, and the project was already behind its original schedule due to the earlier delays in the tendering process.

5. Background of Collecting Research Information

In the beginning of the project, we had three different systems for collecting research information: TUTCRIS, SoleCRIS, and Julki. TUTCRIS (Elsevier Pure) was used as the basis of the new common system, TUNICRIS. TUTCRIS had already gone through one data migration as all the data from the former research register (established in 1989) had been migrated into Pure in 2015. SoleCRIS (maintained by a Finnish company called Solenovo) was used by the former University of Tampere. SoleCRIS was originally adopted in 2008, which was the first year the publication data were collected through the system and validated entirely in the library. Before 2008, publication data collection at University of Tampere had not been centrally organized, and the information and metadata of the publications were quite poor. All former publications were still included in SoleCRIS. The data of publications before 2008 were the poorest. The third system was Julki, used by the Pirkanmaa Hospital District. Julki was also maintained by a Finnish company, and it is used only in hospital districts in Finland. Julki was not a full-fledged CRIS, as it only included metadata about the publications. TUTCRIS and SoleCRIS were both CRIS platforms that also included researcher profiles and research activities.

In Finland, universities have reported their publication metadata to the Ministry of Education and Culture since 2011 [36]. The data collection includes basic bibliographic

information about the publications and is based on different publication types. Publications from university hospitals also became part of this data collection in 2015 [37]. However, university hospitals had a slightly different classification system for publications. The Pirkanmaa Hospital District also seemed not to have performed the validation of publications as thoroughly as the universities.

The Ministry of Education and Culture has had its own instructions for data collection over the years. There have been changes over the years, for example, in the definitions of peer review and open access. The data model has also been expanded over the years and new information has been added, for instance, open access information and corporate collaboration. In the common TUNICRIS system, the metadata were mainly structured according to the newest instructions, so all the old differing metadata had to be converted into the new model.

In some cases, the metadata in the former system were correct according to the instructions of the Ministry of Education and Culture but were not transferable to Pure as such. For example, in SoleCRIS, there was no difference between an edited journal or an edited book. They were all in the same format, and we had to divide them using the ISSN and ISBN numbers as well as journal names as the separator. Fortunately, there were not many such occurrences. The biggest problem was the personal names, which is discussed in detail in the data wrangling section.

For over 14,000 publications, we had to merge the publication metadata from two or three different systems if the same publication was reported in more than one system. Merging the data using DOIs was quite straightforward, although older publications often lacked DOIs. Additionally, inconsistencies were sometimes observed in the reporting of author and journal names, making it impossible to rely solely on them for merging. In such cases, the merging was based on titles. However, even the titles were not always identical. To maximize the amount of information obtained, we prioritized the information with the greatest number of characters when merging data in a specific field.

The data for research activities (conference presentations, peer review and editorial work, memberships, etc.) had more differences. Research activities have been part of CRIS platforms for years, but there have not been any common instructions or categories on how to collect the data. Information on research activities is not usually validated in the library, so the metadata are not very uniform. Research activities were often reported with minimal metadata. In 2020, a national code list for research activities was published in Finland [38]. This code list is mainly based on the Pure activity model. Consequently, the research activities from SoleCRIS (about 89,000 records) had to be converted to the research types in the Finnish code list. This conversion was manually performed using Excel.

6. Migration of the Legacy Data

In the beginning of the project, we knew that we would not get access to the final data until the project had been running for a few months. Because it was not possible to suspend the registration of new publications for the duration of the migration project, it was necessary to strive for automation and rely on manual modifications of data only in borderline cases.

An early inspection of the records extracted from legacy systems revealed that we needed to build the transformation process using tools which support interactive hands-on approaches working with data. Even though the data appeared to be sound when inspected in year-by-year batches, there had been a myriad of changes in data validation guidelines over the lifespan of legacy systems, so, as a whole, the data were highly heterogeneous. Therefore, instead of creating conventional transformation scripts designed to run non-interactively, we decided to build our data transformation process using an interactive computational notebook approach typically employed in data science [39].

Using RStudio Notebooks as a workspace, it was possible to run small sections of transformation, inspect the result, and iteratively improve individual sections of the code. RMarkdown syntax used in the notebooks made it convenient to write out the thought

process and take notes, because in RMarkdown documents, executable snippets of scripts can be smoothly intertwined with documentation. Using RStudio as the workspace for transformation also made it easy to quickly create diagnostic graphics and visually inspect subsets of data.

All steps prior to the creation of the final XML import files for Pure were executed in RStudio, except for the manual modification tasks which were distributed to the project team and handled using familiar spreadsheet tools. Data ready for XML transformation were exported as CSV tables and imported to the BaseX XML database. Using an XML database for creating XML import files made it possible to write conversion scripts using XQuery, which is a highly efficient programming language especially designed for querying and transforming structured document data [40].

The development of XQuery scripts for generating XML files for data import was carried out in BaseX GUI, which offers many options for working with XML data interactively, including full-fledged visualization functionalities [41].

The following R packages and Python libraries were indispensable in data wrangling.

Tidyverse is a collection of R packages for common tasks in data analysis projects, regardless of the domain of the analysis. Functionalities include data import, tidying the data into a consistent form, data transformations such as adding new variables and joining data, and filtering data tables. The guiding design principle of all tidyverse packages is human centeredness, ease of use, and learnability. The goal is to support human data analysts, who are not commonly professional programmers, in their tasks [42].

Reclin2 is a R package which provides a toolkit for record linkage and deduplication based on inexact keys [43]. Record linkage involves identifying and connecting records that are related to the same entity across different data sources. It uses statistical measures to calculate the probability that two records from different datasets refer to the same entity, based on record attributes which are not necessarily unique and exact.

The caret package simplifies the creation of regression and classification models in R [44]. It offers user-friendly functions for every step of the process: from splitting and pre-processing the dataset, through training and tuning the model, to evaluating its performance. In our case, we used caret to train a decision tree classification model using Quinlan's [45] C5.0 algorithm. The model was used to impute missing data in one of the required fields.

Probablepeople is a Python library for parsing unstructured person and organization names into their components and labeling each component [46]. Parsing is implemented using conditional random fields, a statistical modeling method commonly used in natural language processing. Out of the box, Probablepeople is able to recognize common Western names, but by adding context-specific training data, it is possible to improve the accuracy of parsing. In our case, we added training data for Finnish, Swedish, and Romanized Asian names.

6.1. Identified Subproblems in Data Wrangling

Splitting the data wrangling to individual subproblems made it easier to mix and match different problem-solving approaches and tools. It also helped us to avoid bottlenecks, as many of the subproblems were independent of each other, and it was possible to make progress even if problem solving in one area was found to be more difficult and time-consuming than anticipated.

Identified subproblems were:

- Structuring author bylines;
- Matching affiliated authors with byline authors;
- Enriching publication records using data from Scopus and WoS;
- Imputing missing data;
- Resetting known bad data;
- Merging of duplicate publication records;
- Merging of author information of duplicated publication records.

6.1.1. Structuring Author Bylines

Export files obtained from two of the legacy systems had author information in a format where all publication authors were saved as unstructured text typically following the byline formatting of the publication. On the other hand, our target system required each author to be saved individually, with surname and first name separated from each other. The Pure XML schema has an option for storing unstructured author names, but according to the vendor, using this option was not possible in the data import.

Tokenizing bylines to individual authors proved to be the first major hurdle of the migration project. Simply by visually inspecting the data, it became clear that while there were author bylines which should be easy to tokenize using uniform delimiter characters, there were plenty of cases where the delimiter character varied or the delimiter used between authors was also used in-between author name parts. There were also differences on how to write the names, for example:

1. First name Last name
2. Last name, F.
3. Last name First name
4. Last name, First name

Publications with lengthy bylines had been truncated, so that not every author had been necessarily mentioned if the author was not affiliated to the parent organizations. In the Julki system, it was common to use the three first authors of the publication and then the authors of the host organization. In SoleCRIS, the first 20 authors were reported as well as the authors of the host organization. In addition to person names, bylines contained organization names, for example, names of research groups.

To find the best possible strategy for splitting up bylines to individual authors and organization names, we needed to algorithmically determine the delimiter character used in-between authors and name parts. To do this, we computed diagnostic features for each byline, including semicolon, whitespace, and comma character count in bylines, word count, total character count, and the ratios of these basic diagnostic numbers, such as word-to-semicolon or comma-to-semicolon ratio. Inspecting the relationship of diagnostic variables and the actual data, it was possible to develop a strategy for splitting the byline data to smaller subsets where most bylines in one subset shared a common delimiter character. Subsets containing bylines which were deemed to be too difficult to computationally tokenize were manually processed, for example, bylines with a common delimiter between authors and name parts.

The tokenized byline data consisted of the full names of authors and organization names. To structure names further into their components and to distinguish organization names from person names, data was processed using the Probablepeople Python library. The resulting data were far from perfect and required manual cleaning, but without Probablepeople or another similar computational solution, the structuring of the bylines would not have been accomplished.

6.1.2. Matching Affiliated Authors with Byline Authors

Export files which had affiliated author information available saved every author individually, together with a personnel registry identifier and the name of the organization unit which the author was affiliated to. Most of the publications had affiliated author names also saved to the author byline, but in a substantial number of publications, affiliated authors had not been saved to the author byline, even if this should have been done according to the guidelines.

Matching publication authors to affiliated authors was performed in multiple iterations. As previously stated, the quality of the data was not uniform, and therefore a strategy was adopted where the first matching iteration was performed using the author's last name and first name, the second round using the author's last name and first initial, and the third round using an algorithmically generated matching key.

Because manual checking for all computationally structured names would have been too time-consuming, a method for matching names with mixed-up surnames, first names, and first initials was needed. Interchanged surnames and first names still contain the same constituent parts, individual letters, so by joining the first name and last name parts of the names, lowercasing them, removing non-alphabetic letters and diacritics, and finally ordering characters alphabetically, it was possible to construct a key for matching names using the Jaro–Winkler similarity. The Jaro–Winkler similarity belongs to a category of algorithms called string distance metrics, which measure the similarity or dissimilarity between two character strings [47].

Because matching was performed for a single publication at a time, the risk of matching two non-related names was reduced. We decided that the risk of name collision was worth taking, because alternatively, there might have been hundreds, if not thousands, of publications where a single author might have had multiple representations in the final author information section.

6.1.3. Enriching Publication Records with Scopus and WoS Data

To make it easier to conduct bibliographic analyses in the future, we decided to augment the publication data using WoS and Scopus identifiers and DOI. Technically, it would also have been possible to use other metadata from WoS and Scopus, such as author information, but because it was not clear whether we could legally use metadata loaded from Scopus and WoS en masse, we decided to include only identifiers to avoid possible copyright infringement. Another major and laborious challenge with Scopus and WoS author data would have been the manual validation of all internal authors.

We exported the data from WoS and Scopus in Bibtex format, and using the R bibliometrix library, the data were read into an R dataframe [48]. From the loaded data we only used DOIs, Scopus, and WoS identifiers and publication titles. DOI was used as a matching key for publications where DOI was available. For publications without DOI, fuzzy matching based on the Jaro–Winkler similarity was used to match publications using publication and host titles such as journal title as an inexact matching key.

6.1.4. Imputing Missing Data

Approximately 18,500 records from one of the legacy databases were missing the Finnish Ministry of Education publication type classification. Therefore, we needed a way to automatically classify them. We trained a model in caret using data from one of the source legacy databases which had publication type classification information available. Our intuition was that even though the legacy databases had been distinct, and practices on how information had been recorded had varied, all organizations had followed the same guiding instructions to record the data. Guidelines state, for example, which fields are mandatory for a given publication type. We reasoned that it would be likely that data from all our legacy databases exhibit similar publication-type-dependent patterns, and data from one database could be used to train a model for imputing missing data in other legacy data sources.

For training the model, we created derivative features which indicate the presence of metadata fields in the data. It should be noted that we did not use the actual values saved to the metadata field, but a binary value indicating whether a particular field contained any data. In total, the model was trained using 15 features derived from fields such as DOI, ISBN, parent publication title, and volume.

The model was built using Quinlan's [45] C5.0—decision tree algorithm and validated using ten-fold cross-validation. In ten-fold cross-validation, the dataset is split to ten equal parts, and the model is trained using nine of these folds and tested using data of the one remaining fold. The process is repeated ten times, every time using a different part of the dataset for testing [49]. We achieved 0.88 accuracy with the model, which was a sufficiently good starting point for manually checking the robustness of the classification result.

For imputing missing language code information, we used cld2 (Google's Compact Language Detector 2) to detect the language of the publication using a combination of article title and journal and/or series name as an input.

6.1.5. Resetting Known Bad Data

Since the guidelines of publication data collection had changed over the years, we used year-based thresholds for resetting fields with known dubious-quality data prior to the given year threshold. Therefore, it was possible to include at least basic data from records made in earlier years even if some of the data had quality problems.

6.1.6. Merging of Duplicate Publication Records

To merge duplicate publication records, we used the reclin2 R package to create a lookup table of all matching combinations of the source database records. The merged records were constructed field by field for records where data from multiple sources was available.

For solving cases where values in fields were in disagreement, a set of rules was constructed to identify the likely correct value. We decided to prioritize values from legacy sources whose quality and potential problems were known to the project team via prior experience.

In some cases, disagreement between values was rooted in differing practices in recording publication information. We wanted to avoid throwing away usable data, so fields such as subject keywords and the fields of science classification were merged instead of choosing data just from one source.

6.1.7. Merging of Author Information of Duplicated Publication Records

In merging author information of duplicated publication records, the strategy was to first match the affiliated authors at the publication level, because it was known that a substantial number of authors are likely to be double- or triple-affiliated in our data. Matching was performed using reclin2, utilizing first name and surname as matching keys and publication identifier as a blocking variable. In this stage, we had already merged publication records, so it was known which publication records in legacy databases were instances of the same publication. Using this information, it was possible to perform the author matching at the publication level, instead of matching every author record to each other in all legacy databases. This made the matching process more efficient and reduced the possibility of false matches. After the merging of matching affiliated authors, the remaining non-matching affiliated authors were combined to a single set on a record-by-record basis.

Because lengthy bylines had been truncated in the data, it was not possible just to use byline data from one legacy database record. In many publications with a lengthy byline having tens of authors, only the first few authors had been recorded in addition to affiliated authors, regardless of their position in the byline. Therefore, it was possible that all three legacy databases had the same publication saved with differing byline information, and to get best possible result, byline authors from all legacy databases had to be merged to a single set at the publication level and deduplicated using reclin2.

7. Assessing the Quality of Data Migration

As a whole, the data migration was successful, but different users have different standards for data quality. People who spend the most time using Pure are editors who validate the data submitted by researchers, and bibliometricians who use the data in publication analyses and other reporting tasks. From an editor's point of view, the data migration was sufficiently successful, and day-to-day operations were not significantly impacted by the deficiencies in the data. Nevertheless, certain irregularities in the new master data continue to cause minor frustrations. For example, the presence of duplicate

journal records has led to some confusion. These problems are also sometimes encountered by end users, i.e., researchers, when they enter new data into the system.

Another problem faced by editors and end users concerned duplicate publication records. The merging of duplicate records is made quite simple in Pure, but if neither record has complete and correct author information, the process can be quite slow. After the migration and initial clean-ups, between 1500 and 2000 duplicate publication records remained in the system (approximately 2% of the total number of publications). These have been manually merged after the migration as time permits.

For bibliometric analyses, the initial data quality was not acceptable. For the purpose of research evaluation, data quality is essential, especially when individual researchers are being evaluated. Because some publications were not linked to corresponding researcher profiles in the new instance of Pure, the Tampere University Library metrics team continues to manually validate all researcher profiles that are included in bibliometric analyses. Even when limiting the task to publications within the designated year range used in analyses, it has so far required days of manual verification. Many of the corrections made have required using external information sources to determine if a given publication should be linked to a certain researcher profile and, as such, would have been extremely difficult to perform automatically.

Negative feedback from end users has been relatively rare. Only a few researchers have contacted the system administrators to notify them of issues related to their profile. One reason for this might be the fact that the newer publication data were in better condition due to better source data. Also, usually there are not many errors related to individual profiles, which makes it harder to spot these, especially if a researcher has an extensive publication history.

8. Discussion

Dealing with a massive dataset originating from several different source systems was a challenging situation to begin with. In the early stages of the project, the project team thought that cleaning and structuring the author data could be manually performed in a spreadsheet using regular expressions. This created a false impression of data regularity. However, as the project progressed, it became clear that the data were highly heterogeneous and could not be handled by any conventional means. For example, the separator used between personal names could change in the middle of the author list, or the same characters were used as a separator between different authors and an individual author's first and last names. Manually structuring this type of data in a spreadsheet would have been an impossible task. In total, approximately 335,000 occurrences of personal names were processed in the migration, with data correction manually performed in a spreadsheet for about 10,000 cases. As problems with the data were revealed one by one (i.e., as the project's complexity and scope increased), the project schedule was not realistically revised.

Some of the issues in source data that caused a great deal of additional work only became clear in the final stages of the project, when the records from different source databases were merged in October 2020. Problems arose, for example, when the same publication had different author information in different source databases. For example, only a few of the first authors might have been recorded, or the author information might have been given based on the organization, regardless of that author's order among other authors. There was also variation in the spelling of names between source databases. This significantly complicated the merging of source data, and the final result achieved was not entirely satisfactory in terms of quality.

Some of the problems encountered in handling author data could have been avoided by using data downloaded from reference databases. This would have been possible according to the Finnish Ministry of Education and Culture reporting guidelines, as author data can be reported in the form in which it appears in reference databases. Using author data from reference databases could have saved time in the migration process, as the biggest problems and the need for manual correction were related to the handling of author

data. On the other hand, the internal author data would also have needed to be validated by librarians.

Another major challenge in the project was the allocation of resources. Since we were not initially prepared to handle the legacy data migration ourselves, we had not allocated a data specialist with programming skills to the project. In hindsight, it would have been critical to reassess the resources and the entire project schedule when this change became apparent. On the other hand, we had no choice but to proceed with the resources we had, because the system provider declined to carry out the migration. This kind of expertise is difficult to hire from outside because it requires highly specialized skills in data wrangling and also context-specific knowledge of collecting research information.

The key lesson of the project is that a realistic assessment of the initial situation provides a good basis for estimates of workload, resource allocation, and scheduling. In this case, the source data were not known well enough at the start of the project. On the other hand, some problems became apparent only when the data were loaded into Pure, making them difficult to anticipate and prepare for in advance. Furthermore, the project manager had practically three different roles during the project: running the project, being a technical expert in the project group, and also maintaining one of the existing systems (TUTCRIS). For future projects of this scale, having a dedicated project manager and the required resources for data wrangling from the start are absolute necessities. This would ensure that the project receives dedicated attention, allowing for better coordination, resource management, and technical expertise.

9. Plans for Further Development

Elsevier Pure has been actively developed since the implementation project, and we now have new tools available to address the remaining data issues. Perhaps the most significant new opportunities are provided by the new Pure API with write access, which enables us to edit data in the system much more efficiently than previously.

Until now, our efforts to clean the data after the implementation project have almost exclusively focused on publication and person data, due to the fact that person data is directly related to the end user experience, and publication data is frequently reported on and analyzed. However, the landscape of research evaluation is evolving, and the importance of other kinds of research output and activities is rising. For instance, the Agreement on Performing Research Assessment by the Coalition for Advancing Research Assessment [50] explicitly calls for the recognition of outputs beyond journal articles. In addition, since 2022, it has been possible to export research activity data into the Finnish national research portal. Therefore, it will also be important to clean data related to research activities and other types of research output.

As we deal with more diverse data and more complex integrations in the future, the importance of persistent identifiers (PID) will keep on growing. If all publications had DOIs or other PIDs and all researchers had ORCIDs, migrations like the one described in this article would be much simpler. By making sure that PIDs are used whenever possible, we can proactively mitigate the challenges associated with future research system migrations, minimizing potential difficulties.

10. Conclusions

This article described a complex CRIS implementation project involving the migration of around 120,000 legacy publication records from three different systems. The large amount and highly heterogeneous quality of the data made it necessary for the project team to adopt innovative methods and tools for automated data processing. Machine learning methods and various data wrangling tools were used to automatically structure author data, fill missing data, and merge data from different sources. Although the project was significantly delayed from its original plan, the project eventually succeeded in accomplishing its goals. Notably, the entire data migration was carried out within the library, without hiring any outside help. At the time of writing, the Tampere University Research Portal

(https://researchportal.tuni.fi, accessed on 21 August 2023) exhibited an extensive collection of over 145,000 publications and 113,000 activities, covering the entire publication history of the former Tampere universities and the Pirkanmaa Hospital District. The project was also a significant learning experience that improved the team's readiness to further process and refine publication data in the future.

As research organizations continue to procure and develop new systems, new data migrations keep occurring periodically. The significance of data quality, international standards, and persistent identifiers becomes increasingly important, particularly when information is collected on a national or international level. This is further accentuated by the development of national CRIS platforms and other research aggregators. To successfully handle these migrations, it is essential to have robust platforms, tools, and methods in place to mitigate potential data quality issues that may arise during the process.

Author Contributions: Conceptualization, Y.L., M.L., T.H., J.N. and T.L.; software, M.L.; investigation, Y.L., M.L., T.H., J.N. and T.L.; writing—original draft preparation, Y.L., M.L., T.H., J.N. and T.L.; writing—review and editing, Y.L., M.L., T.H., J.N. and T.L.; project administration, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. OKM. Korkeakouluille uusi Rahoitusmalli. *Opetus- ja Kulttuuriministeriö.* 2019. Available online: https://okm.fi/-/korkeakouluille-uusi-rahoitusmalli (accessed on 2 July 2023).
2. Pölönen, J.; Pylvänäinen, E.; Aspara, J.; Puuska, H.M.; Rinne, R. Publication Forum 2010–2020: Self-evaluation report of the Finnish quality classification system of peer-reviewed publication channels. *Web Publ. Fed. Finn. Learn. Soc.* **2021**, *9*. [CrossRef]
3. Puuska, H.M. The Research Information Hub as an Access Point to Finnish Research [PowerPoint Slides]. EuroCRIS Spring 2019 Membership Meeting. 2019. Available online: https://hdl.handle.net/11366/986 (accessed on 9 May 2023).
4. Laitinen, S.; Sutela, P.; Tirronen, K. Development of Current Research Information Systems in Finland. In Proceedings of the 5th Conference on Current Research Information Systems (CRIS 2000), Helsinki, Finland, 25–27 May 2000. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2db2be9d764dfbbd31e75220936e5682c05f9193 (accessed on 6 June 2023).
5. Asserson, A.; Jeffery, K.G. Current Research Information Systems (CRIS): Past, Present and Future. *Wissenschaftsmanagement* **2009**, *15*, 41–44. Available online: https://hdl.handle.net/1956/6929 (accessed on 7 June 2023).
6. Joint, N. Current research information systems, open access repositories and libraries: ANTAEUS. *Libr. Rev.* **2008**, *57*, 570–575. [CrossRef]
7. Velásquez-Durán, A.; Ramírez Montoya, M.S. Research Management Systems: Systematic Mapping of Literature (2007–2017). *Int. J. Adv. Sci. Eng. Inf. Technol.* **2018**, *8*, 44–55. Available online: https://hdl.handle.net/11285/632426 (accessed on 9 May 2023). [CrossRef]
8. Pinto, C.S.; Simões, C.; Amaral, L. CERIF–Is the standard helping to improve CRIS? *Procedia Comput. Sci.* **2014**, *33*, 80–85. [CrossRef]
9. Houssos, N.; Jörg, B.; Dvořák, J.; Príncipe, P.; Rodrigues, E.; Manghi, P.; Elbæk, M.K. OpenAIRE guidelines for CRIS managers: Supporting interoperability of open research information through established standards. *Procedia Comput. Sci.* **2014**, *33*, 33–38. [CrossRef]
10. Príncipe, P.; Rettberg, N.; Rodrigues, E.; Elbæk, M.K.; Schirrwagen, J.; Houssos, N.; Holm Nielsen, L.; Jörg, B. OpenAIRE guidelines: Supporting interoperability for literature repositories, data archives and CRIS. *Procedia Comput. Sci.* **2014**, *33*, 92–94. [CrossRef]
11. De Castro, P.; Schirrwagen, J.; Karaiskos, D.; Dvořák, J.; Bollini, A.; Bonis, V.; Gasparis, N.; Tsoukala, V.; Manghi, P.; Príncipe, P. Progress in the Implementation of the OpenAIRE Guidelines for CRIS Managers. *Procedia Comput. Sci.* **2017**, *106*, 104–111. [CrossRef]
12. De Castro, P.; Dvořák, J.; Simons, E. OpenAIRE compatibility for CRIS systems: Recent progress. *Procedia Comput. Sci.* **2019**, *146*, 182–189. [CrossRef]
13. Rybinski, H.; Skonieczny, L.; Koperwas, J.; Struk, W.; Stepniak, J.; Kubrak, W. Integrating IR with CRIS–a novel researcher-centric approach. *Program* **2017**, *51*, 298–321. [CrossRef]

14. Simons, E.; Jetten, M.; Messelink, M.; van Berchum, M.; Schoonbrood, H.; Wittenberg, M. The Important Role of CRIS's for Registering and Archiving Research Data. The RDS-project at Radboud University (the Netherlands) in Cooperation with Data-archive DANS. *Procedia Comput. Sci.* **2017**, *106*, 321–328. [CrossRef]
15. Schöpfel, J.; Prost, H.; Rebouillat, V. Research data in current research information systems. *Procedia Comput. Sci.* **2017**, *106*, 305–320. [CrossRef]
16. Jippes, A.; Steinhoff, W.; Dijk, E. NARCIS: Research information services on a national scale. In Proceedings of the 5th International Conference on Open Repositories (OR2010), Madrid, Spain, 6–9 July 2010. [CrossRef]
17. Wenaas, L.; Karlstrøm, N.; Vatnan, T. From a national CRIS along the road to Green Open Access–and back again: Building infrastructure from CRIStin to Institutional Repositories in Norway. In Proceedings of the 11th International Conference on Current Research Information Systems, Prague, Czech Republic, 6–9 June 2012. Available online: https://hdl.handle.net/11366/116 (accessed on 9 May 2023).
18. Ibanez, K.H. Research Portal Denmark: An Update [PowerPoint Slides]. EuroCRIS Strategic Membership Meeting 2022. 2022. Available online: https://hdl.handle.net/11366/2251 (accessed on 9 May 2023).
19. Palavesm, K.; Joorel, J.S. IRINS: Implementing a Research Information Management System in Indian Higher Education Institutions. *Procedia Comput. Sci.* **2022**, *211*, 238–245. [CrossRef]
20. Lopes, A.L. Integrating a local CRIS with the PTCRIS synchronization ecosystem. *Procedia Comput. Sci.* **2019**, *146*, 166–172. [CrossRef]
21. Kaliuzhna, N.; Auhunas, S. Research Information Infrastructure in Ukraine: First steps towards building a national CRIS. *Procedia Comput. Sci.* **2022**, *211*, 230–237. [CrossRef]
22. Pinto, A.L.; de Carvalho Segundo, W.L.R.; Dias, T.M.R.; Silva, V.S.; Gomes, J.C.; Quoniam, L. Brazil Developing Current Research Information Systems (BrCRIS) as data sources for studies of research. *Iberoam. J. Sci. Meas. Commun.* **2022**, *2*. [CrossRef]
23. Kremenjaš, D.; Udovičić, P.; Orel, O. Adapting CERIF for a national CRIS: A case study. In Proceedings of the 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 28 September–2 October 2020; IEEE: Piscataway, NJ, USA; pp. 1633–1638. [CrossRef]
24. Chudlarský, T.; Dvořák, J. A national CRIS infrastructure as the cornerstone of transparency in the research domain. In Proceedings of the 11th International Conference on Current Research Information Systems, Prague, Czech Republic, 6–9 June 2012. Available online: https://hdl.handle.net/11366/95 (accessed on 7 June 2023).
25. Tuglular, T.; Gurdal, G.; Can, G.K.; Ozdemirden, A.S. Repository Landscape in Turkiye and GCRIS: The first National Research Information System. *Procedia Comput. Sci.* **2022**, *211*, 222–229. [CrossRef]
26. EuroCRIS. CRISCROS—A Working Group to Bring Together National and Regional CRIS Initiatives. 2022. Available online: https://eurocris.org/criscros-%E2%80%93-working-group-bring-together-national-and-regional-cris-initiatives (accessed on 25 October 2023).
27. Azeroual, O.; Saake, G.; Abuosba, M. Data quality measures and data cleansing for research information systems. *J. Digit. Inf. Manag.* **2018**, *16*, 12–21. Available online: https://hdl.handle.net/11366/632 (accessed on 25 October 2023).
28. Azeroual, O.; Schöpfel, J. Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries. *Publications* **2019**, *7*, 14. [CrossRef]
29. Azeroual, O.; Saake, G.; Abuosba, M.; Schöpfel, J. Solving problems of research information heterogeneity during integration—Using the European CERIF and German RCD standards as examples. *Inf. Serv. Use* **2019**, *39*, 105–122. [CrossRef]
30. Van den Berghe, S.; van Gaeveren, K. Data quality assessment and improvement: A Vrije Universiteit Brussel case study. *Procedia Comput. Sci.* **2017**, *106*, 32–38. [CrossRef]
31. Buchmayer, C.; Greil, M.; Hikl, A.L.; Kaiser-Dolidze, O.; Miniberger, C. Usability on the Edge: The Implementation of u:cris at the University of Vienna. *Procedia Comput. Sci.* **2014**, *33*, 103–109. [CrossRef]
32. Siciliano, L.; Schmidt, S.; Kinzler, M. BoRIS and BIA: CRIS and institutional repository integration at the Free University of Bozen-Bolzano. *Procedia Comput. Sci.* **2014**, *33*, 68–73. [CrossRef]
33. Grenz, D.; Lery, T.; Ward, M.; Mastoraki, E.; Baessa, M. A CRIS in the Desert: The Implementation of Pure at KAUST A Case Study in Information Exchange. *Procedia Comput. Sci.* **2017**, *106*, 176–182. [CrossRef]
34. McDonnell, R.; Kerridge, S. Research information management system (KRIMSON) at Kent. *Procedia Comput. Sci.* **2017**, *106*, 160–167. [CrossRef]
35. Bevan, S.; Harrington, J. Managing research publications: Lessons learned from the implementation of a Current Research Information System. *Serials* **2011**, *24*, 26–30. [CrossRef]
36. Ilva, J. Juuli—Yliopistojen Julkaisutiedot Yhteiseen Käyttöliittymään. *Tietolinja 1/2013*. 2013. Available online: http://urn.fi/URN:NBN:fi-fe201306043828 (accessed on 8 May 2023).
37. Poropudas, O. Tutkimuslaitosten Ja Yliopistollisten Sairaaloiden Liittyminen OKM- Julkaisutiedonkeruuseen [PowerPoint Slides]. Opetus- Ja Kulttuuriministeriö. 2015. Available online: https://slideplayer.fi/slide/2918504 (accessed on 8 May 2023).
38. CSC. Koodistot: Activity and Prize Types and Roles. 2020. Available online: https://koodistot.suomi.fi/codescheme;registryCode=research;schemeCode=aktiviteetitjarooli (accessed on 8 May 2023).
39. Shen, H. Interactive notebooks: Sharing the code. *Nature* **2014**, *515*, 152. [CrossRef]
40. Anderson, C.B.; Wicentowski, J.C. *XQuery for Humanists*; Texas A&M University Press: College Station, TX, USA, 2020.

41. Grün, C.; Gath, S.; Holupirek, A.; Scholl, M.H. XQuery Full Text Implementation in BaseX. In *Database and XML Technologies*; Bellahsène, Z., Hunt, E., Rys, M., Unland, R., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2009; p. 5679. [CrossRef]
42. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the Tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [CrossRef]
43. Van der Laan, D.J. reclin2: A toolkit for record linkage and deduplication. *R J.* **2022**, *14*, 320–328. Available online: https://rjournal.github.io/articles/RJ-2022-038 (accessed on 15 April 2023). [CrossRef]
44. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
45. Quinlan, J.R. *C4. 5: Programs for Machine Learning*; Morgan Kauffmann Publishers: San Francisco, CA, USA, 1993.
46. Deng, C.; Gregg, F. Probablepeople Documentation. 2015. Available online: https://probablepeople.readthedocs.io/en/latest (accessed on 6 April 2023).
47. Cohen, W.W.; Ravikumar, P.; Fienberg, S.E. A Comparison of String Distance Metrics for Name-Matching Tasks. In Proceedings of the 2003 International Conference on Information Integration on the Web, Acapulco, Mexico, 9–10 August 2003; pp. 73–78. Available online: http://people.csail.mit.edu/emax/public_html/papers/approximate-string-matching/iiweb03.pdf (accessed on 15 April 2023).
48. Aria, M.; Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *J. Informetr.* **2017**, *11*, 959–975. [CrossRef]
49. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*; Liu, L., Özsu, M.T., Eds.; Springer: Boston, MA, USA, 2009. [CrossRef]
50. Coalition for Advancing Research Assessment. Agreement on Reforming Research Assessment. Available online: https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf (accessed on 20 June 2023).