Automatically Creating Hyperlinks of References in a Digital Library of Scientific Articles

Daniela R. N. de Mello[1], Rafael Dueire Lins[1,2], and Gabriel Pereira e Silva[2]

[1] Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
{drnm, rdl}@cin.ufpe.br
[2] Departamento de Computação, Univ. Federal Rural de Pernambuco, Recife, Brazil
{rafael.dueirelins, gabriel.franca}@ufrpe.br

Abstract. The access to scientific knowledge and the progress of bibliometric analysis depend on the integrity and interoperability of research data. However, a significant portion of historical scientific production remains located in legacy formats, hindering its traceability. This article addresses these barriers by presenting a framework for locating and validating bibliographic references from the proceedings of the International Symposium on Telecommunications (ITS) 2010. The data from 1,775 references located in 127 articles stored in PDF format on CD-ROM were manually extracted and processed using a Python pipeline integrated with the Crossref REST API. Considering the limitations from incorrect or summarized writing of the reference article data, the proposed framework dynamically adjusts the relevance of author, title, and year according to the completeness of each citation, enabling robust validation even in the presence of metadata inconsistencies. A validation decision rule was implemented, in which references were automatically validated only if they achieved a confidence score above 80% combined with a Levenshtein distance tolerance of two character edits. The results demonstrate that the proposed approach achieved an F1-Score of 93.1% and a Precision of 100.0% on a curated ground-truth sample. Although automation accelerates the retrieval of correct URLs, manual curation remains essential to resolve ambiguities. The resulting reference database improves navigability, accessibility, and preservation of the ITS 2010 data collection and provides a reproducible framework for metadata enrichment and hyperlink generation in any digital library.

Keywords: Digital libraries · References · Hiperlink navigation

1 Introduction

Libraries have been of paramount importance throughout history for the preservation of knowledge, beginning with the ancient Mesopotamian clay tablets. Ancient libraries, such as the Ashurbanipal Library, maintained thousands of documents, allowing the survival of literary, religious, and administrative works. During the Middle Ages, monastery libraries in Europe preserved ancient Greek

and Roman knowledge through manuscript copying. The Library of Alexandria and the House of Wisdom in Baghdad functioned as vital hubs for scholars, scientists, and researchers to study in various fields. Libraries were and are still centers of cultural and religious significance. Beginning in the 17th and 18th centuries, the rise of subscription and public libraries meant access to information was no longer restricted solely to wealthy patrons, scholars, or monks. Academic libraries became integral to educating future generations and supporting academic communities.

Throughout history, libraries have served as repositories, transforming from restricted, elite archives to public institutions dedicated to education and knowledge access. The Internet is the most democratic of libraries in human history. Today, anyone with access to the WWW has unlimited access to data in all areas, but the quality of such data accessed should be carefully checked, in order to provide reliable information.

The proceedings of technical events provide a written record of the development of a research area. However, past proceedings are often restricted to those who participated in the event, making them difficult to obtain and to access; in some cases, records may even disappear, creating gaps in the history of the evolution of events and even research areas. In Brazil, the first event to have its proceedings in digital format was the 1996 Congress of the Brazilian Computing Society (SBC), organized by the second author of this paper back in 1996. Such proceedings were offset printed and handed to all participants. Besides that, the proceedings were also on CD, a novel technology, which were distributed to approximately 10% of the participants. That initiative may be seen as a pioneering step in the generation of a digital library of proceedings. The authors sent the printed version of their papers, that were hand collected to be sent to be printed. To generate the CD of the proceedings, all the papers were scanned (true-color, 200 d.p.i) and generated a multi-page pdf-document file, which were indexed by title and authors making possible to navigate on the CD. The development of the Nabuco Project [6] [7] provided the correct "feeling" for which scanning resolution to adopt. The 200 d.p.i. resolution was proved adequate only in the recent paper [10]. That initiative was so *avant garde* that only in 1999 the proceedings of the SBC Congress was released on CD, a medium that had already become popular.

In 1997, the second author of this paper was the co-organizer of the Congress of the Brazilian Telecommunications Society (SBrT), and **all** the participants received the proceedings both on offset printed and CD versions. This time, some authors have already sent their contribution in pdf, while others still sent theirs articles in paper, which were 200 dpi scanned in true-color and saved in pdf.

A decade later, in 2007, the second author decided to innovate again, releasing a DVD with all the proceedings from 1997 to 2007. For that, the 1998 to 2001 had to be scanned, as there were only printed versions of the proceedings. Then, the LiveMemory Project [8] [9] was born. The most important challenge then was to fit all the proceedings of eleven years in a DVD. Again, paper proceedings

were scanned page by page in true-color 200 dpi. To save space and fit in a DVD, all pages had to be binarized.

Fig. 1. DVD covers of SBrT 2008 (left) and SBrT 2009 (right) handed to each of the participants of The Brazilian Telecommunications Symposium in 2008 and 2009

Only pages that had graphics or figures were automatically de-assembled, had the text part binarized and the figures and graphs converted into grayscale (one should remember that the proceedings were offset printed in grayscale). That strategy allowed for "visual continuity" whenever flipping the pages of the papers in the digital library of the proceedings. The participants still received a printed version of the proceedings.

In 2010, the second author of this paper was the chairman on the International Telecommunications Symposium, a joint quadrennial event between the IEEE and the Brazilian Telecommunications Society. A new challenge was set, to provide each participant with a DVD with all the history of the SBTs and ITSs. The cover of such a DVD is presented in Figure 2. The Sociedade Brasileira de Telecomunicações is possibly the first scientific society to have all its proceedings in a digital library. The ITS 2010 proceedings were only in DVD format, saving a lot of money in the organization of the event.

With the advent and the widespread use of the Internet, in 2011 the SBrT decided to move all the content of the ITS 2010 DVD into a web-site, and update it yearly with the proceedings of the events, making their organization even

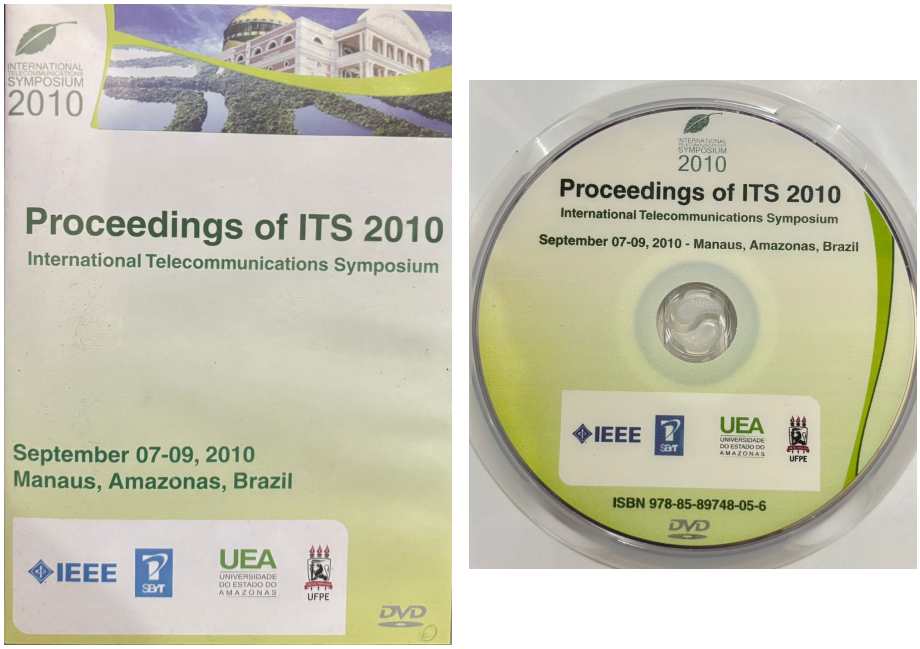

Fig. 2. DVD cover and the DVD provided to each participant of ITS 2010, a digital library with all the papers in the proceedings of the symposia of the Brazilian Telecommunications Society.

cheaper as the proceedings are only available in the WWW. Today, most scientific societies have the proceedings of their events and journals on the Internet, available for their members and subscribers.

A WWW digital library can be defined as a specialized repository of multimedia assets, including: text, audio, and video, backed by systematic frameworks for content curation, structural organization, and information retrieval [12]. Despite its potential, a significant portion of the scientific output of the late 20th and early 21st centuries lack structured metadata and are often incompatible with current bibliographic management systems. Hyperlinking became a solution to connect web pages and resources. However, as the amount of digital content grew exponentially, maintaining reliable and persistent links became a significant challenge. URLs (Uniform Resource Locators) were commonly used to reference web pages, but they were prone to situations where URLs became invalid due to changes in web addresses or the disappearance of resources. Recognizing the need for a robust and persistent linking system, the International DOI Foundation (IDF) [4] was established in 1998. The IDF collaborated with various stakeholders, including publishers, libraries, and technology providers, to develop the DOI system. Many editorial houses assigned a DOI to all their publications, allowing to uniquely identify them. However, many scientific associations did not take that important measure.

The navigation in the SBrT digital library of papers, likewise all other repository of academic papers, could be done by searching the author name in the whole library or by the title of the article. It is important to note that the scanned articles yielded pdf-image files. Thus, the SBrT library is a mixture of pdf-image and pdf-text files. This work focuses on the aspect of finding the list of the "References" in each paper in the library and automatically obtaining and validating the hyperlinks of references in a digital library of scientific articles, allowing inter-navigation among documents.

Generating metadata from pdf-image files is frequently compromised by font encoding errors, OCR (Optical Character Recognition) artifacts, or non-standard citation styles with abbreviated or incomplete terms. Furthermore, manual cataloging of extensive bibliographies is time-consuming and prone to human error, which makes the development of automated or hybrid validation frameworks necessary. This article proposes a robust methodology for enriching and validating bibliographic metadata from digital libraries. The proceedings of the International Symposium on Telecommunications (ITS) 2010, which represents a critical body of scientific knowledge that, although stored digitally on physical media, faces significant barriers to large-scale accessibility due to format limitations and the publication period of the articles. By integrating manual extraction with a custom Python-based pipeline, the Crossref REST API was used to retrieve metadata from references belonging to 127 articles in the ITS 2010 proceedings. The approach focuses on an adaptive validation mechanism that uses Levenshtein distance [5] and a weighted confidence scoring system to handle incomplete or corrupted citations.

Unlike traditional string-matching techniques, the defined method implements a threshold requiring a weighted score greater than 80% and a Levenshtein distance lower than or equal to two character edits for automatic acceptance, ensuring that the resulting dataset meets a high standard of academic precision. This work not only contributes to the digital preservation of the ITS 2010 collection but also provides a replicable framework for researchers dealing with the recovery of legacy scientific data, being completely compatible to be used in any digital library in the Internet. If the reference item in a paper matches with a paper found in the Intermet site of a publisher (ACM, IEEE, IAPR, etc) it has a hyperlink added to it. If the DOI is also present for the matching reference, then it is included, updating the references in the original article. Several challenges are faced here:

shortened information: Sometimes for space limitations the list of authors is reduced to the first author and *et al*, some other times the title of the cited article is reduced with *etc*. It is common to abbreviate "Transactions" by "Trans.", etc.

wrong or missing information: Sometimes authors' names are misspelled or incomplete. The year of publication is missing or wrong, etc.

Although several Natural Language Processing (NLP) tools and standard reference extraction models currently exist, they are predominantly based on

modern, standardized metadata structures and well-formatted native digital documents. The critical gap addressed in this work is the handling of legacy scientific proceedings, processed by OCR, in which citations are commonly unstructured, suffer from coding noise, and frequently contain incomplete author lists (e.g., truncated with 'et al.'). Our main contribution is a specialized and adaptive framework that not only retrieves metadata simply but also actively validates it against historical anomalies. By integrating a weighted confidence scoring mechanism with Levenshtein distance, the approach presented here dynamically adjusts to partial metadata, providing the researcher with an initial result regarding the reliability of the information found, a challenge in which standard extraction tools natively fail. Previous research in record linkage and entity resolution has demonstrated the effectiveness of similarity metrics and confidence-based matching strategies for identifying equivalent records across heterogeneous datasets. The use of weighted similarity functions for handling incomplete metadata is discussed [1], while Levenshtein (1966) introduced the edit-distance metric widely adopted for approximate string matching. The present work adapts these principles to the context of legacy scientific proceedings, where OCR artifacts, incomplete citations, and non-standard reference formats create additional challenges for reliable metadata validation.

The structure of this paper is organized as follows: Section 2 details the extraction and pre-processing of the dataset and describes the implementation of the similarity metrics and the automated classification system; Section 3 presents the results, followed by the conclusions obtained.

2 Materials and Methods

This section further describes the source material used in this research together with the automated metadata retrieval methodology adopted in this paper.

2.1 Data Collection and Pre-processing

The *International Telecommunications Symposium (ITS) 2010* DVD includes the proceedings of the SBrT events from 1983 to 2010. The primary source of data used in this research consists of 1.775 references extracted from 127 articles in the proceedings of the last year, 2010. To ensure the analysis reflects the documents exactly as they were provided in the original media, a manual transcription was performed following a literally "as-is" protocol.

Each reference was recorded in a structured spreadsheet containing Title, Author, and Year information, maintaining the original text provided in the PDFs, regardless of potential human errors or incomplete metadata. This preservation of the raw data state is essential for ensuring the automated validation process is tested against real-world data conditions, accounting for variations in author styles and citation practices present in the original proceedings. This methodological choice ensures that the study evaluates the effectiveness of locating the reference links based on the material actually distributed to the scientific community.

[0001] 1.28-Tbit/s DWDM Transmission on a Recirculating Loop Using DQPSK and FBG Dispersion Compensation			
ID	Author	Title	Year
[1]	C. Kalmanek	Next generation photonic network directions	2008
[2]	M. H. Eiselt & B. T. Teipen	Requirements for 100-Gb/s metro networks	2009
[3]	H. Sun et al.	Modulation formats for 100Gb/s coherent optical systems	2009
[4]	D. Liu et al.	Requirement of modulation bandwidth for 100Gb/s optical DQPSK transmission using one dual-drive Mach-Zehnder modulator	2008
[5]	P. J. Winzer & R-J Essiambre	Advanced optical modulation formats	2006
[6]	P. J. Winzer & R-J Essiambre	Advanced optical modulation formats	2008
[7]	P. J. Winzer et al.	100-Gb/s DQPSK transmission: from laboratory experiments to field trials	2008

Fig. 3. Sample of bibliographic references extracted from source article [0001].

2.2 Automated Metadata Retrieval via Crossref API

To evaluate the extracted metadata, an automated retrieval pipeline was implemented in Python using the Crossref REST API. Crossref operates as an interconnected "Research Nexus" that goes beyond merely assigning persistent identifiers to digital content; it functions by building a rich, open network of relationships among diverse research entities such as journal articles, datasets, software, grants, and peer reviews. Crucially, Crossref does not operate as an automated web crawler. Instead, its foundational infrastructure relies entirely on the active and continuous deposition of metadata by its member institutions, such as publishers and universities [3]. The search process is systematically executed for each parsed reference through the following stages:

1. **Query Formulation:** The original metadata components extracted from the PDF—specifically the author string, document title, and publication year—are stripped of trailing whitespaces and concatenated into a unified bibliographic query string. For each reference, a bibliographic query string Q was generated by concatenating the original author (A_{orig}), title (T_{orig}), and year (Y_{orig}):
$$Q = \{A_{orig} \cup T_{orig} \cup Y_{orig}\} \tag{1}$$
2. **API Request and Polite Pool Access:** An HTTP GET request is dispatched to the Crossref works endpoint (`https://api.crossref.org/works`). To ensure reliability and avoid server timeouts, the script identifies itself via the `User-Agent` header with an academic email address. This practice routes the requests through Crossref's *Polite Pool*, granting them priority processing [3]. The search utilizes the `query.bibliographic` parameter, and the response is strictly limited to the top match (`rows=1`) to optimize network bandwidth and computational efficiency.
3. **JSON Parsing and Data Extraction:** Upon receiving a successful response (HTTP 200), the system parses the returned JSON payload to extract the best-matching item. The algorithm isolates the retrieved title, iterates through the author list to concatenate their family names, and evaluates multiple date fields (such as *published-print*, *published-online*, and *issued*) to identify the most accurate publication year. The API returns a set of

items, from which the top-ranked candidate's metadata $(T_{ret}, A_{ret}, Y_{ret})$ is extracted for similarity analysis.

4. **Rate Limiting:** To comply with Crossref's fair use policies and prevent temporary IP address bans due to excessive traffic, a deliberate throttling mechanism is enforced. The algorithm pauses for 0.4 seconds between consecutive requests, establishing a sustainable retrieval rate.

2.3 Similarity Metrics and Scoring Functions

Distinct metrics were established to quantify the reliability of the retrieved metadata across different bibliographic fields.

Title Similarity via Levenshtein Distance To account for character errors inherent in the PDF metadata extraction process, the normalized Levenshtein distance was applied, a metric based on the foundational work focused on correcting deletions, insertions, and substitutions in information sequences [5]. The title similarity (S_L) using LD is defined as:

$$S_L = \left(1 - \frac{L(T_{orig}, T_{ret})}{\max(|T_{orig}|, |T_{ret}|)}\right) \times 100 \tag{2}$$

Where L denotes the minimum number of single-character edits required to change the original extracted title (T_{orig}) into the retrieved title (T_{ret}).

Year Similarity The year similarity metric (S_Y) evaluates the temporal match between the extracted publication year (Y_{orig}) and the API's returned date (Y_{ret}). Recognizing that recorded publication dates can occasionally fluctuate between print and online versions or face extraction delays. A step-function penalty based on the absolute difference in years was applied:

$$S_Y = \begin{cases} 100 \text{ if } |Y_{orig} - Y_{ret}| = 0 \\ 80 \quad \text{if } |Y_{orig} - Y_{ret}| = 1 \\ 50 \quad \text{if } |Y_{orig} - Y_{ret}| = 2 \\ 0 \quad \text{otherwise} \end{cases} \tag{3}$$

2.4 Final Confidence and Decision Model

The weighted score (C_w), is the definitive measure used to validate the metadata retrieved. This index utilizes the Levenshtein Similarity (S_L) to ensure robustness against character-level noise, combined with the sequence matching ratio for the authors (S_A) and the year penalty score (S_Y).

A heuristic weighting system was developed to adapt to different citation styles. The calculation is performed through an adaptive logic that distinguishes two primary scenarios based on the citation structure, specifically the presence of the *"et al."* string, which indicates incomplete author metadata:

– **Standard Metadata:** In cases where the citation is presumed to be complete, the confidence score is calculated as an unweighted arithmetic mean. The title similarity (S_L), author similarity (S_A), and year similarity (S_Y) contribute equally to the final validation:

$$C_w = \frac{S_L + S_A + S_Y}{3} \tag{4}$$

– **Partial Metadata (Presence of *et al.*):** To mitigate the impact of incomplete author lists and avoid false negatives, the system adaptively redistributes the weights. The validation is shifted primarily to the title (70%) and the publication year (25%), while the author metadata is assigned a minimal weight (5%):

$$C_w = (0.70 \cdot S_L) + (0.05 \cdot S_A) + (0.25 \cdot S_Y) \tag{5}$$

2.5 Classification

The final confidence determination (C_w) combines macro-level textual agreement with character-level similarity by integrate the weighs and the Levenshtein similarity metric (S_L). Previous studies in record linkage and entity resolution indicate that no universally optimal similarity threshold exists; instead, thresholds must be calibrated according to the characteristics of the dataset and the desired balance between precision and recall [1]. Therefore, the threshold adopted in this work ($C_w > 80\%$) was empirically selected to prioritize precision and reduce the likelihood of false-positive matches. Additionally, an Levenshtein distance tolerance of $L \leq 2$ was introduced to accommodate minor OCR and typographical errors while maintaining strict matching criteria, following the general principles of approximate string matching and record linkage [1]. The classification logic is formally defined as:

$$\text{Status} = \begin{cases} \text{Validated if } (C_w > 80\%) \wedge (L \leq 2) \\ \text{Review} \quad \text{if } (50\% \leq C_w \leq 80\%) \vee (L > 2 \text{ with } C_w > 80\%) \\ \text{Rejected} \quad \text{if } C_w < 50\% \end{cases} \tag{6}$$

This methodology ensures that entries are automatically accepted only when they exhibit both a strong macro-textual weighted alignment ($C_w > 80\%$) and a highly accurate character-level title match ($L \leq 2$). This strict associative approach effectively mitigates the risks of false positives from legacy font artifacts, cleanly routing references with significant OCR discrepancies to the manual review pipeline.

This methodology ensures that only entries with a Levenshtein distance lower than or equal to two character edits and high weighted alignment are automatically accepted, effectively mitigating the risks associated with the legacy font artifacts identified in the original proceedings.

2.6 Evaluation Measures and Validation Protocol

Due to the extensive volume of the dataset (1.775 references), a comprehensive manual verification of all retrieved URLs was computationally and practically unfeasible. To rigorously evaluate the performance of the proposed algorithm and establish a ground truth, a random statistical sample was extracted for manual curation. To determine the exact sample size (n) required for the known, finite population of $N = 1775$ references, the sampling estimation techniques [2] were applied. The study parameters were established assuming maximum statistical variance ($p = 0.5$), a standard 95% confidence level ($Z = 1.96$), and a target margin of error of $d = 0.095$ (9.5%). The initial sample size approximation for an infinite population (n_0) is calculated as:

$$n_0 = \frac{Z^2 \cdot p(1-p)}{d^2} \approx 106.41 \tag{7}$$

Because the dataset population is finite, this initial value is adjusted using the finite population correction (FPC) formula:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \approx 100.44 \tag{8}$$

Applying that standard rounding principles yields a final curated sample of $n = 100$. Such a sample size effectively balances statistical validity with the inherent operational constraints of human relevance assessments, which is a standard practice in Information Retrieval evaluation methodologies [11]. During the curation phase, the URLs and DOIs automatically retrieved by the Crossref API were manually accessed and verified against the original extracted text to confirm their accuracy. The performance of the algorithm was then quantified using standard Information Retrieval measures: Precision, Recall, and the F1-Score. Precision measures of the system exactness, the percentage of automatically validated URLs that were indeed correct, while the Recall measures its completeness, the system ability to identify all correct URLs without erroneously sending them to manual review.

3 Results and Discussion

The data processing revealed that the primary articles exhibit approximately an average of 13.82 references per document. Furthermore, a significant variation in reference volume was observed, with articles containing a range between a minimum of 4 and a maximum of 45 citations. This considerable breadth indicates a high degree of diversity in the theoretical depth of the documents analyzed.

3.1 Algorithm Performance on the Sampled Ground-Truth

Due to the large volume of the test set (1,775 references), a comprehensive manual verification of all retrieved URLs was unfeasible. A random sample of 100

references was extracted for manual curation, considering a ground-truth establishment. The evaluation measures (Precision, Recall, and F1-Score) reported in this study reflect the behavior of the proposed algorithm on this statistically representative subset. The manual curation of the random sample (n=100) demonstrated that the proposed algorithm operates with a Precision of 100% and a Recall of 87.1%, culminating in an F1-Score of 93.1%. The perfect Precision score underscores the high reliability of the classification system: no incorrect URLs were automatically validated by the algorithm. The slightly lower Recall is a direct and intended consequence of the strict validation threshold established in Eq. 6 (Confidence $> 80\%$ and Levenshtein Distance ≤ 2). In the context of digital heritage preservation, avoiding False Positives—which would silently corrupt the scientific repository with incorrect hyperlinks—is prioritized over maximizing automated matches. This justifies the conservative approach of the algorithm, cleanly allocating uncertain references to the manual review pipeline while guaranteeing the absolute accuracy for the validated entries.

3.2 Quality and Data Validation (CrossRef)

The effectiveness of metadata retrieval and the integrity of the references were measured using the weighted confidence index and the Levenshtein distance:

- **Validation Rate (Validated):** 43.1% (756 references) were validated with high confidence (similarity above 80% or an exact Title match). This demonstrates that the majority of the original dataset contains precise metadata that is easily traceable in official databases.
- **Review Zone (Review):** 30.4% (533 references) showed similarity between 50% and 80%. These cases typically arise from minor discrepancies in author abbreviations, conference titles, or a one-year variance between online and print publication dates.
- **Rejection Rate (Rejected):** 26.6% (466 references) scored below 50%. This subset identifies references that may contain severe typos, missing titles, or consist of "gray literature" (technical reports and manuals) not formally indexed on CrossRef.

3.3 Ablation Study

To directly address the individual contribution of each metadata component to the predictive power of the algorithm, an ablation study was conducted using the curated ground-truth sample ($n = 100$). The objective is to demonstrate that standard isolated matching techniques are insufficient for legacy datasets, justifying the need for the proposed adaptive weighting system. The classification performance was evaluated by isolating the similarity measures. In each ablated scenario, the confidence score (C_w) was derived from a single parameter, bypassing the adaptive heuristic:

- **Title-Only:** Classification relying exclusively on the Levenshtein distance similarity (S_L).

- **Author-Only:** Classification relying exclusively on the author string matching ratio (S_A).
- **Year-Only:** Classification relying exclusively on the temporal penalty score (S_Y).
- **Full Proposed System:** The complete algorithm, integrating S_L, S_A, and S_Y with the adaptive weighting mechanism triggered by incomplete metadata markers (e.g., "et al.").

Table 1 presents the comparative Information Retrieval measures for each configuration. The ablation results indicate a significant drop in the overall F1-Score

Table 1. Ablation study evaluating the contribution of individual metadata components on the curated sample ($n = 100$).

Model Configuration	Precision	Recall	F1-Score
Title-Only Matching	100.0%	65.7%	79.3%
Author-Only Matching	100.0%	4.3%	8.2%
Year-Only Matching	83.6%	80.0%	81.8%
Full Proposed System	**100.0%**	**87.1%**	**93.1%**

when the algorithm relies on isolated components. The Author-Only model exhibits a severe drop in Recall (4.3%), which is consistent with the highly variable nature of legacy OCR extraction where author lists are frequently truncated. Conversely, the Year-Only baseline sacrifices the exactness of the proposed scheme, dropping Precision to 83.6% and introducing undesirable False Positives. While the Title-Only model performs reasonably well by maintaining perfect Precision, its Recall remains limited (65.7%). It is the integration of the adaptive weighting system—which dynamically adjusts the influence of S_L, S_A, and S_Y when encountering incomplete lists—that successfully boosts Recall to 87.1% while sustaining the maximum Precision of 100.0%. This confirms that integrating all components with a structural awareness of the citation format is essential for processing historical scientific archives.

3.4 Observed Bibliographic Variations

During the execution of the validation pipeline and the final association phase, several characteristics inherent to the source material were observed. Since the dataset maintained the original writing from the PDFs, the following scenarios were identified and addressed by the proposed system:

- **Incomplete References:** A subset of citations provided by the authors lacked standard metadata, such as missing author names or incomplete titles, common in legacy paper submissions.
- **Transcription and Encoding Noise:** Instances of missing characters or typographical variances resulting from either human error during the original paper finalization or technical artifacts from the PDF font layers.

- **Non-standard Citation Format:** Variations in citation formats were observed, with entries alternating between "name, surname," "initial, surname," and "surname, name," in addition to differences in the punctuation used to separate multiple authors.
- **Non-standard Citation Grouping:** Cases where authors grouped multiple distinct bibliographic works under a single index (e.g., `[1]`), which required manual disaggregation to ensure correct link association.

These observations validate the importance of using robust similarity metrics such as the Levenshtein distance, as the system must navigate these pre-existing irregularities to successfully retrieve the correct hyperlinks to easily access each reference.

A more in-depth analysis of the 26.6% rejection rate reveals specific challenges inherent in extracting legacy data. Considering the present evaluated dataset from ITS 2010, the most common causes for match failures include:

- **Unindexed 'Gray Literature':** Citations referring to internal technical reports, historical manuals, or historical books that are not present in digital databases, and publications that lack representation in the Crossref database.
- **Publications not represented in the Crossref database:** For the Crossref feature to function correctly in locating files, the file must be registered in its database. In some cases, such as technical reports, historical manuals, theses, and dissertations, it has been found that those files were not registered and therefore it is impossible to confidently establish a match.
- **Extreme Metadata Truncation:** Instances where authors provided only a last name and an abbreviated title, leaving insufficient data for the adaptive weighting system to confidently establish a match.
- **References to systems:** Some references are to software, APIs, and systems that do not have published written versions; they only reference the system used in the research.

4 Conclusions

This work presented an adaptive metadata validation framework for automatically locating and validating bibliographic references in legacy scientific digital libraries. The manual cataloging of articles combined with the implementation of the automated validation algorithm enabled the analysis of the following points:

- **Efficiency of Automation:** The script successfully classified and validated approximately 60% of the database autonomously, drastically reducing the manual effort required for bibliographic verification.
- **The importance of a standardized reference structure:** Manual data mapping was necessary due to the lack of standardization in file references, making it difficult for an OCR system to correctly find the articles.

- **Reliability of the Original Base:** Most references (43.1%) have a direct and robust match in global databases, lending credibility to the research corpus.
- **Need for Human Intervention:** The presence of 30.4% of references in a "Review" state indicates the need for targeted manual adjustments, particularly where similarity was affected only by formatting issues.
- **Identification of Gaps**: The rejection rate of 26.6% does not necessarily mean that the articles do not exist; instead, it indicates that a quarter of the database requires in-depth investigation to correct metadata, confirm unindexed sources, and analyze search limitations of the Crossref system.

Some limitations should be considered when interpreting the results. First, the evaluation was conducted on a single collection of telecommunications proceedings, which may limit the generalizability of the findings to other scientific domains. Second, the ground-truth validation was performed on a statistically representative sample rather than on the entire dataset. Finally, the proposed weighting parameters and decision thresholds were empirically selected and may require recalibration when applied to collections with different citation characteristics or OCR quality levels.

These findings provide a foundation for the next stages of research, ensuring that future work, such as automatic citation mapping, uses only verified and accurate data. Technical events and journals should encourage their authors to include the DOI for the cited papers, whenever available.

Acknowledgments. Rafael Dueire Lins holds a Senior Researcher grant from CNPq - Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico (Brazilian Government).

Disclosure of Interests. The authors hereby declare having no competing interests.

References

1. Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science & Business Media.
2. Cochran, W. G. (1977). Sampling Techniques (3rd ed.). John Wiley & Sons.
3. Crossref: REST API Documentation. https://www.crossref.org/documentation/.
4. DOI Foundation. The Identifier: What is a DOI? `https://www.doi.org/the-identifier/what-is-a-doi/`, last visited 03/06/2026.
5. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
6. Dueire Lins, R., Rosa, L.G., França Neto, L., Guimarães Neto, M.S., "An Environment for Processing Images of Historical Documents", Microprocessing & Microprogramming, pp. 111-121, North-Holland, 1994.
7. Dueire Lins, R. Nabuco - Two Decades of Document Processing in Latin America, JUCS - Journal of Universal Computer Science, 17(1): 151-161, 2011. https://lib.jucs.org/issue/4191/

8. Dueire Lins, R., Torreão, G., Silva, G.P.: Content Recognition and Indexing in the LiveMemory Platform. In: Graphics Recognition. Recent Advances and New Challenges. GREC 2009, LNCS, vol. 6020, pp. 253–264. Springer, Heidelberg (2009)
9. Dueire Lins, R., Silva, G.P., Torreão, G., Alves, N.F.: Efficiently Generating Digital Libraries of Proceedings with The LiveMemory Platform. In: 7th International Telecommunications Symposium (ITS 2010). Manaus, Brazil (2010)
10. Dueire Lins, R., Nunes de Mello, D.R. and Correa de Oliveira, R., Which is the most suitable scanner resolution for documents? Detailing the answer given to the question raised by Professor George Nagy. In ACM Symposium on Document Engineering 2024 (DocEng '24), (2024) https://doi.org/10.1145/3685650.3685672
11. Manning, C. D., Raghavan, P., Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
12. Witten, I.H., Bainbridge, D.: How to Build a Digital Library. Morgan Kaufmann, San Francisco (2003)