Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 1069–1074
DOI: 10.15439/2023F1515
ISSN 2300-5963 ACSIS, Vol. 35

EpiDoc Data Matching for Federated Information Retrieval in the Humanities

Sylvia Melzer*†, Meike Klettke‡, Franziska Weise*, Kaja Harter-Uibopuu* and Ralf Möller†
*Universität Hamburg
Centre for the Study of Manuscript Cultures, Warburgstraße 26, 20354 Hamburg, Germany
Email: {sylvia.melzer, franziska.weise, kaja.harter}@uni-hamburg.de
†University of Luebeck
Institute of Information Systems, Ratzeburger Allee 160, 23562 Luebeck, Germany
Email: moeller@uni-luebeck.de
‡University of Regensburg
Faculty for Computer Science and Data Science, University Street 31, 93053 Regensburg, Germany
Email: meike.klettke@ur.de

***Abstract*—The importance of federated information retrieval (FIR) is growing in humanities research. Unlike traditional centralized information retrieval methods, where searches are conducted within a logically centralised collection of documents, FIR treats each information system as an independent source with its own unique characteristics. Searching these systems together as a centralised source results in lower precision in humanities research, even when the research data itself is structured and stored according to standardised guidelines such as EpiDoc, and requires the need to be able to trace the origin of records to avoid incorrect historical conclusions. Matching of queries against all data sets in each source is proving less effective. A global search index that enables traceable matching of key values deemed relevant would provide a more robust solution here. In this article, we propose a solution that introduces a novel EpiDoc data matching procedure, facilitating traceable FIR across distinct epigraphic sources.**

I. Introduction

IN THE field of humanities, the need for federated information retrieval (FIR) is becoming increasingly important [5], [13], [20], [21]. FIR refers to the process of searching for relevant information across distributed and autonomous information systems within a database federation. A database federation provides a logical centralisation of data without the need to change the physical implementation of databases and maintain the identity of autonomous developed databases.

Information systems have emerged in some humanities projects, e.g., the epigraphy projects "Epigraphische Datenbank zum antiken Kleinasien" (EDAK) [23] and "Collection of Greek Ritual Norms" (CGRN) [3], they are not collaboratively searchable because these information systems run in heterogeneous hardware and software environments, and have different data models. Although both projects use an epigraphy-specific XML format called EpiDoc [7], a customized version of TEI (Text Encoding Initiative) [22], is additionally provided to exchange research data. The purpose of the EpiDoc format is to enhance machine-readability, and effectively searching for specific information within EpiDoc files relies on correctly matching the research data extracted from these files across different sources.

The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796.

Fig. 1. Different "date" representations in EpiDoc. Top: EpiDoc file from the CGRN project. Bottom: EpiDoc file from the EDAK project.

Although an EpiDoc schema was developed in each of the two projects, in Fig. 1 it is shown that the general structure is similar, however, in detail the XML tags are applied differently. In an example, according to the TEI guidelines, the XML tag "origDate" is used to represent dates. In practice, concrete date specifications for "origDate" vary. In CGRN a date presents a century and in EDAK an epoch (see Fig. 1). When specifying the place, mapping from both sources becomes even more difficult because the semantics of the place terms are different. While in CGRN the place name is specified in the XML tag "ref", in EDAK the tag "placeName" is used (see Fig. 2). Only an expert in this field can say exactly how the places can be mapped onto each other. For an efficient FIR, a correct matching of data sets from different sources must be defined in advance to provide precise IR results.

In the humanities, mapping of data from different sources is still done manually. The manual procedure is necessary

Thematic track: International Workshop on AI in Digital Humanities, Computational Social Sciences and Economics Research

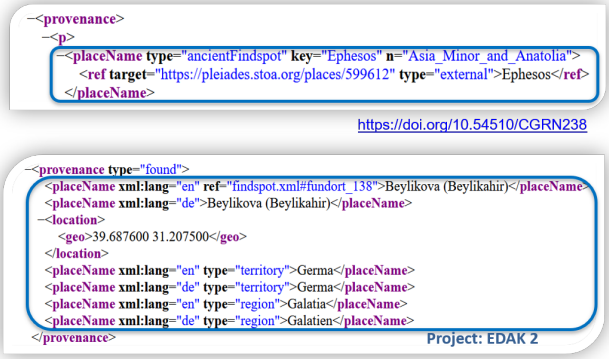

Fig. 2. Two different representations of "placeName" in EpiDoc. Top: EpiDoc file from the CGRN repository. Buttom: EpiDoc file from the EDAK repository.

due to the need to achieve a high degree of precision in semantic evaluation, which is crucial to avoid making false statements about cultural artefacts. The current challenge is to automatically match the data, which is stored in various EpiDoc formats. Searching through hundreds upon thousands of EpiDoc files can be challenging, so it can be advantageous to transfer data represented in EpiDoc to a relational database. The advantages of transforming data from EpiDoc into a relational database are:

- (i) higher query answering performance because relational databases are optimized for query answering,
- (ii) scalability which makes relational databases a good choice for large data sets, and
- (iii) opportunity to use pre-defined data types and to define relationships between data elements.

In the CGRN and EDAK projects, relational mappings exist beside the XML representations. We use both, the existing relational mappings and the XML tags, to compute relevant key values and use these values for our new indexing approach for a precise and performant EpiDoc data matching. If the relational mappings are missing, it is possible to use large language models such as Generative Pre-trained Transformer 2 (GPT-2) [24] or other versions to generate the required relational mappings. Computing the relational mappings is done by entering the specific EpiDoc schemas into the language model, which can then generate the required mappings.

This article presents a novel method to overcome the difficulties of automatically matching epigraphic data from different sources with different EpiDoc schemas while ensuring semantic precision. The proposed method offers potential benefits for humanities scholars by maintaining precision while minimising heterogeneity caused by differences in data semantics. Furthermore, the proposed EpiDoc data matching approach can be applied to an FIR to offer scholars in the humanities access to a wider range of information from distributed and autonomous sources.

The remaining article is structured as follows: Section II gives an overview of some work on the representation of epigraphic data, as well as selected approaches to match these data. Section III describes a new matching process for epigraphic research data to enhance correct semantic mapping. This process provides the basis for enabling FIR, which is described in Section IV. Section V concludes this article and gives an outlook.

II. Related Work

Studies emphasise the need for careful selection of repositories in federated search to avoid describing objects that are not searchable. A prototype federated search engine [17] or cross-domain information system [11] has been developed to address this problem by integrating selected repositories. However, manual mapping is usually required to link different content with high precision. This article proposes an automatic data mapping that eliminates the need for manual mapping.

Data in EpiDoc.XML is syntactically represented as XML documents. Existing work on XML matching can be applied for schema matching and mapping. These methods typically begin with element-level similarity assessment [1] and then extend to data set level comparison [10]. For comparing EpiDoc data sets, this article suggests adopting similarity functions used for XML elements, such as Levenshtein distance [14] or the Soundex algorithm [9]. The overall similarity between data sets can be evaluated using metrics like Jaccard distance [8].

Applications processing semi-structured data often require schema matching for tasks like schema integration and schema clustering. Previous research, such as [10] and [18], has defined similarity functions for semi-structured data (DTDs) and schema fragments (from XSD) to address these needs. In our approach, we can focus specifically on matching different variants of EpiDoc, assuming that all input data are in this format. This eliminates the need to rely on general schema matching algorithms and allows us to treat the different EpiDoc variants as dialects of the same language.

III. Matching of Epigraphic Research Data

Matching data sets involves the process of comparing two or more data sets to identify similar elements. In the following the general process of matching data and the process of matching EpiDoc data are presented (cf. [4] and see Fig. 3).

A. General Process of Matching Data

a) Data pre-processing: The initial step in data pre-processing involves preparing the data sets for matching. This involves cleaning, formatting, and standardizing the data sets to ensure compatibility and effective comparison. This may also include removing duplicates and identifying missing data.

b) Indexing: The complexity of matching records increases with the number of records to be matched. Indexing is a strategy to pre-select potential matches and leads to a reduction in the number of matches. Indexing usually involves identifying the key variables that will be used to match the data sets. These variables may include unique identifiers, such as the titles of editions, dates, or places. A traditional indexing

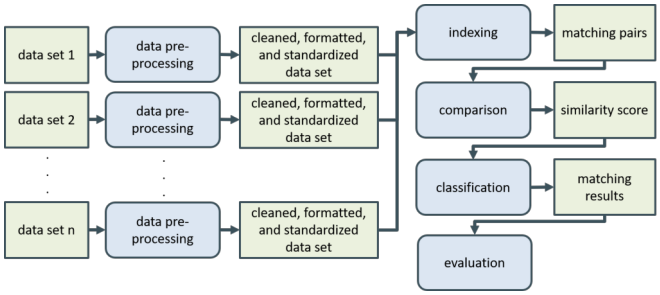

Fig. 3. The general process of matching n data sets. Based on [4] (extended)

method has the name blocking [2]. The method compares only those records that are based on the same so-called blocking criteria. In this article, we present an advanced blocking method that considers both the XML and relational structure of databases created based on the XML files to determine the matching criteria and then identify the candidate pairs for comparison.

c) Comparison: In this step, the data sets are compared to identify matches. For this purpose, the similarity between the key values used during indexing, but also additional other values, will be calculated.

d) Classification: Based on the data set comparison, the matching records are classified as match or non-match.

e) Evaluation: The final step involves validating the matched data sets and reviewing the results for accuracy and completeness. This may involve checking for errors, inconsistencies, or missing data and making any necessary adjustments.

B. Process of Matching EpiDoc Data

1) Data Pre-processing: As a starting point, we use the EpiDoc files from the projects EDAK and CGRN. The EpiDoc data are transformed into a relational database using the so-called databasing on demand (DBoD) approach [12], [19]. The DBoD approach is used for building project-specific information systems on demand in a few hours and with few resources. The DBoD process consists of the following steps:

1) Transformation of all EpiDoc files into one CSV file. The mapping of the EpiDoc XML elements into a canonical mapping was carried out via the widely used EpiDoc XSLT stylesheets [6] as defined by epigraphers (cf. data representations on the websites: [3], [23]).
2) Insert all the data from the CSV file in a database instance.

From the website it could be inferred that both projects have different relational representations. The EDAK project has the column names “Edition”, “Inscription type”, “Obejct type”, “Region”, “Place”, “Date (epoch)”, “Text”, . . ., while CGRN has the column names “Edition”, “Provenance”, “Place”, “Date (century)”, “Text”, ...

To ensure comparability between centuries and epochs, we have transformed them into a starting time (notBefore) and an ending time (notAfter). For instance, the “imperial” epoch was converted to notBefore=1 and notAfter=300. This conversion allows for a standardized representation of time periods, facilitating analysis and comparison across different historical eras.

In [4] it is suggested to standardize tables, so that the relational models consist of the same entity types, and thus facilitates the data matching process. Since in our case the mapping rules are given (defined by epigraphers), it is not necessary to use the same database models for all databases, especially since in practice this form of database entry would not be accepted by the users due to the diverse requirements. In the indexing step, it is shown that mapping rules, rather than standardisation of databases, are sufficient to successfully perform data matching. If the structure of the relational database is already known, mapping EpiDoc data sets directly to a database instance using XSLT or a similar transformation language may be easier and more efficient. However, if these mapping rules are missing, our hypothesis is that large language models can be useful to define such mapping rules if the input data are the EpiDoc schemas and the EpiDoc guidelines. We tested our hypothesis with the use of ChatGPT [16], which has statistical knowledge, acquired through its extensive training on large data sets. We gave ChatGPT the EpiDoc guidelines as input and the two EpiDoc schemas, and asked it to provide us with the mapping for transforming EpiDoc data into a relational database. As output, we received tables that correspond to the hierarchical structure of the XML file. In total, ChatGPT delivered more tables than needed, but all contents remain taken into consideration.

2) Indexing: Indexing includes identifying the key variables for an efficient schema matching process. For existing relational databases, it can be assumed that the column names belong to the key variables and are used for their project-specific analysis. Therefore, the column names are regarded as key variables. Since matching all key variables is inefficient, a blocking procedure is traditionally used [4]. This reduces the number of comparisons and improves performance. This article employs an alternative approach to existing blocking methods by utilizing the XML schema to identify matching candidates. For seen data (mapping rules known), XSLT is used, while ChatGPT is used for unseen data (mapping rules are not known). The identification of matching candidates is described in the next paragraph.

If two sets A and B of XML tags, where the sets A and B are from different EpiDoc schemes, are mapped to the same element, then that element is a matching candidate to be added to the matching candidate set C. Let $A = \{a_1, \ldots a_i\}$ and $B = \{b_1, \ldots b_j\}$ be sets of XML tags, and let f be a function which represents a mapping from A to B: $f : A \rightarrow B$, then the matching candidates C are given by:

$$C = \{b \in B : \exists a \in A \text{ with } f(a) = b\} \tag{1}$$

In the given example, the EpiDoc schema “EDAK” belongs to set A and “CGRN” to set B. Table I displays the column names used in the respective projects and the

TABLE I
OVERVIEW OF MATCHING CANDIDATES DERIVED FROM THE EPIDOC SCHEMES AND RELATIONAL REPRESENTATION OF THE EPIDOC CONTENT

EpiDoc Schema	**Column name**	**XML tag**	**matching candidate C**
EDAK	Edition	title	no
EDAK	Inscription type	term	no
EDAK	Object type	objectType	no
EDAK	Region	placeName	yes
EDAK	Place	placeName	yes
EDAK	Date (epoch)	origDate	yes
EDAK	Text	div	yes
CGRN	Edition	idno	no
CGRN	Provenance	placeName	yes
CGRN	Date (century)	origDate	yes
CGRN	Text	div	yes

TABLE II
MATCHING DATA OF EDAK AND CGRN

project	**ID**	**placeName**	**Sndx-PN**
EDAK	a_1	Pisidia	P230
EDAK	a_1	Antiochia	A532
EDAK	a_2	Ephesus	E120
CGRN	b_1	Ephesos	E120
CGRN	b_2	Tomis	T520
CGRN	b_3	Athens	A352

corresponding XML tags. The matching candidates are $C = \{\text{placeName}, \text{origDate}, \text{div}\}$.

a) Matching: The data also includes words that sound similar and can also be judged semantically similar, e.g. "Ephesus" and "Ephehos." In the simple comparison, the terms are evaluated as different, even though they are the same place.

The commonly used phonetic coding algorithm Soundex is employed to find a match despite minor differences. Each word is coded into a letter and a three-digit number sequence, words with the same coding are scored as similar. That means for our example that "Ephesus" coded as E120 and "Ephehos" coded as E120 are semantically similar. For more details of the Soundex algorithm is given in [15].

For data represented with the data type "Text" or "Date", the Soundex procedure is not applied as it maps to strings and numeric values. Although "Date" is also a numeric value, it represents a period of time, and the comparison of time periods differs from that of strings and numeric values. This is not considered in the indexing procedure, but in our comparison step.

Using the matching criteria Soundex for the matching candidate "placeName," then the indices and record pairs are presented as shown in Table II. The matching key values P230, A532, and E120 are identified. The only record pair that was identified is (a_2, b_1) for E120.

Formally, the set of matching pairs are computed as follows:

$$f_{\text{Sndx}} : C_{A_{\text{Sndx}}} \rightarrow C_{B_{\text{Sndx}}}$$

$$P = \{(a \in C_{A_{\text{Sndx}}}, b \in C_{B_{\text{Sndx}}}) : \exists a \in C_{A_{\text{Sndx}}} \text{ with } f_{\text{Sndx}}(a) = b\} \quad (2)$$

When comparing the EDAK data set with the CGRN data set, there is only the overlap with one region. This result was to be expected, as it is common in the humanities for research to be conducted in a very specialised area, and it can be assumed that there is little overlap. Nevertheless, in the humanities one is interested in finding other interpretations of texts or even the same data sets. Our next application example shows that our algorithm can also handle larger data sets and that more similar data sets are to be expected in the context of a humanities project. We split one part of the EDAK data set into two so that we have 199 entries in one data set and 201 in the other. Of these, 159 matching candidates were identified based on the same region. The two comparisons (CGRN + EDAK; EDAK part 1 + EDAK part 2) provide the results, as expected, with a high degree of precision.

If comparison is made with the tables or columns provided by ChatGPT, the difference is that there is a larger number of tables or table columns that would have to be compared with each others. As a result, as the number of tables grows, more key candidates emerge and more matches are made than would be necessary. However, if there are no mapping rules, this approach is still helpful because precise results can still be shown.

3) Comparison: The comparison process in schema matching indicates the degree of similarity between two record pairs to determine whether they are a match or not.

The comparison function $c(a_i, b_j)$ maps the matching pairs values of a_i and b_j as well as the pairs with the data type "Text" to a similarity score in the range $[0, 1]$, where 0 indicates no similarity and 1 indicates a perfect match. The comparison function can be defined using different similarity metrics, such as the Jaccard coefficient, cosine similarity, or edit distance, depending on the characteristics of the schema elements and the matching criteria.

In this article, the Jaccard similarity is used to compare the sets of matching terms from P associated with a_i and b_j defined as follows:

$$c(a_i, b_j) = \left| \frac{P(a_i) \cap P(b_j)}{P(a_i) \cup P(b_j)} \right| \quad (3)$$

where $|.|$ denotes the cardinality of a set. For dates, the similarity is computed in the following way. Assuming d_1 and d_2 represent the time periods d_{a_1} to d_{a_2} and d_{b_1} to d_{b_2}, the date similarity is given by:

$$sim_{\text{date}}(a_1, b_1) = \begin{cases} 1 & d_{a_1} \leq d_{b_2} \text{ and } d_{b_1} \leq d_{a_2} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

If there is a temporal overlap, the similarity score is 1, otherwise 0.

For "placeName", the similarity score sim_{place} is also 1 if there is a match, otherwise 0.

TABLE III
COMPARISON

ID	placeName	Date	div	sim_{all}
a_1	Ephesus	0001 - 0300	[text]	
b_2	Ephesos	-550 - 500	[text]	
	1	1	0.5	2.5

The comparison function can be used to rank the candidate matches based on their similarity scores:

$$\begin{aligned}\text{sim}_{\text{all}}(a_i, b_j) = &\, sim_{\text{place}}(a_i, b_j) \\ &+ sim_{\text{date}}(a_i, b_j) \\ &+ c(a_i, b_j)\end{aligned} \tag{5}$$

and to select the best match(es) according to a given threshold or ranking criteria.

The comparison between the two EDAK data sets revealed a similarity score ranging between 2.00 and 2.25. The determination of whether this score indicates a match is explained in the following step.

4) Classification: Classifying the compared record pairs based on their summed similarities is a two-class (binary) classification task. Each compared record pair is classified to be either a match (1) or a non-match (0) depending on a threshold value θ.

The classification of each compared record pair can be based on either the full comparison vectors or on the summed similarities. Based on the summed similarity score, a match is defined as:

$$\text{match} = \begin{cases} 1 & \text{sim} \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

In the context of the project, a good value for θ is between the "number of attributes" divided by 2 and the total "number of attributes" to achieve matching results between approximately 50% and 100%. Formally:

$$\frac{\text{number of attributes}}{2} \leq \theta \leq \text{number of attributes}. \tag{7}$$

If $\theta =$ "number of attributes" (100% similarity), then it could indicate a duplicate. It is important to note that, in general, the optimal value for θ may depend on the specific characteristics of the data sets being compared and the desired level of similarity between them. Now the matching results can be evaluated by experts.

5) Evaluation: In the EDAK project some dates are marked as "Unknown". In this case, the date should have been cleaned up in the pre-processing step or should have been taken into account in the algorithm with another θ value. After we adjusted θ, the similarity score increased.

IV. FEDERATED INFORMATION RETRIEVAL

FIR is an approach used in information retrieval (IR) systems where multiple autonomous and distributed information sources are integrated to provide a unified and comprehensive search experience. Unlike traditional centralized IR methods, which rely on a single collection of documents, FIR treats each information system as an independent source with its own unique characteristics.

```
<xsl:template name="edak-body-structure">
  <!--Date-->
  <xsl:if test="//t:origin/t:p[1]/t:origDate/text()">
    <xsl:value-of select="//t:origin/t:p[1]/t:origDate/@notBefore"/>
    <xsl:value-of select="'|'"/>
    <xsl:value-of select="//t:origin/t:p[1]/t:origDate/@notAfter"/>
    <xsl:value-of select="'|'"/>
  </xsl:if>

  <!--Provenance-->
  <xsl:if test="//t:provenance/t:p/t:placeName/t:ref/text()">
    <xsl:value-of select="//t:provenance/t:p/t:placeName/@key"/>
  </xsl:if>
  <xsl:value-of select="'|'"/>

  <!--text-->
  <xsl:variable name="edtxt">
    <xsl:apply-templates select="//t:div[@type='edition']"/>
  </xsl:variable>
  <xsl:apply-templates select="$edtxt" mode="sqbrackets"/>
  <xsl:value-of select="'|'"/>
</xsl:template>
```

Fig. 4. XSLT code shows that the EpiDoc data is read from the project CGRN

A federation of database systems is called federated database system (FDBS) and integrates multiple autonomous database systems into a single database system. However, the identity of the individual databases is not lost in the merging process. In general, the constituent the physically decentralized databases are interconnected via computer networks.

We have implemented a prototype to integrate the new indexing procedure into an FDBS and thus enable FIR. For this purpose, we have written a Python script combined with the XSLT stylesheets for each project (EDAK and CGRN) that first transforms the EpiDoc data into a relational database model. To do this, it was necessary to adapt the existing XSLT stylesheets to transfer the project-specific mappings from XML to a relational database. Fig. 4 shows the representative XSLT source code representing how the EpiDoc data (date, provenance, and text) is read from the CGRN project.

A further Python script was written for the presented matching process of epigraphic research data to compute the similarity score for the three selected attributes: "Place", "Date" (splitted into "notBefore" and "notAfter" to represent a period), "Text". The result of our script is a tabular listing of all selected attributes that are ranked according to the similarity score. CGRN and EDAK did not provide any results for the area for the place name "Ephesus" and the period 1-300 (notBefore-notAfter). This result was almost to be expected, since research in the field of the humanities is designed in such a way that the projects are usually distinct in terms of content and detail.

As evidence of the applicability of the new data matching process, we generated two data sets from the EDAK project and compared them with each other. In Table IV the top 10 results with the highest similarity are presented. In sum, we have an overall similarity of 39.75% within the result set, which was to be expected.

TABLE IV
TOP 10 RESULTS OF THE EPIDOC DATA MATCHING PROCESS

Edak 1	Edak 2	notBefore	notAfter	Sndx-PN	Sim_Score
TAM V 2, 1152	TAM V 2, 868	1	300	L356	2,25
TAM V 2, 1152	TAM V 2, 1151	1	300	L356	2,21
TAM V 2, 1075	TAM V 2, 1151	1	300	L356	2,20
TAM V 2, 1024	TAM V 2, 987	1	300	G430	2,19
MAMA VII, Nr. 67	Robinson, TAPhA 57 (1926) , Nr. 7	1	300	L250	2,19
Laminger-Pascher, Inschriften Lykaoniens (1992) , Nr. 303	Robinson, TAPhA 57 (1926) , Nr. 7	1	300	L250	2,19
TAM V 2, 1075	TAM V 2, 868	1	300	L356	2,19
TAM V 2, 1026	TAM V 2, 987	1	300	G430	2,17
TAM V 2, 1085	TAM V 2, 1061	1	300	G430	2,17
TAM V 2, 1024	TAM V 2, 1061	1	300	G430	2,17
MAMA VIII, Nr. 315	Robinson, TAPhA 57 (1926) , Nr. 7	1	300	L250	2,17
TAM V 2, 1182	TAM V 2, 868	1	300	L356	2,16

The current implementation so far only supports a search by the three selected categories (attributes). We have also prototypically implemented our new EpiDoc data matching method as an FIR in such a way that a user can enter the attributes "placeName", "notBefore", "notAfter", and "text." As a result, the user receives a list with the matching candidates and the similarity score such as presented in Table IV.

V. SUMMARY AND OUTLOOK

This article is about the increasing importance of federated information retrieval (FIR) in the field of humanities. An FIR, unlike traditional centralised information retrieval methods, treats each information system as an autonomous resource with unique properties. The article proposes a novel EpiDoc data matching procedure which uses the XML schema representations and relational representations to identify the matching candidates, so that on the one hand the number of data matches is reduced and on the other hand the precision is maintained. The new procedure was successfully implemented as a prototype and will be evaluated in the future by transferring it to the productive system at the Universität Hamburg.

REFERENCES

[1] Algergawy, A., Nayak, R., Saake, G.: Element similarity measures in XML schema matching. Inf. Sci. **180**(24), 4975–4998 (2010)

[2] Baxter, R., Christen, P., Churches, T.: A Comparison of Fast Blocking Methods for Record Linkage. Workshop on Data Cleaning, Record Linkage and Object Consolidation at the Ninth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington DC (2003)

[3] Carbon, J.M., Peels-Matthey, S., Pirenne-Delforge, V.: Collection of Greek Ritual Norms (CGRN) (2017-, consulted on 10/05/2023). https://doi.org/https://doi.org/10.54510/CGRN0, http://cgrn.ulg.ac.be

[4] Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Publishing Company, Incorporated (2012)

[5] Demeester, T., Nguyen, D., Trieschnigg, D., Develder, C., Hiemstra, D.: Snippet-Based Relevance Predictions for Federated Web Search. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval. pp. 697–700. Springer Berlin Heidelberg (2013)

[6] Elliott, T., Au, Z., Bodard, G., Cayless, H., Lanz, C., Lawrence, F., Vanderbilt, S., Viglianti, R., et al.: EpiDoc Reference Stylesheets (version 9). Available: https://sourceforge.net/p/epidoc/wiki/Stylesheets/ ((2008-2017)), accessed January 22, 2022

[7] Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S., et al.: EpiDoc Guidelines: Ancient documents in TEI XML (Version 9). Available: https://epidoc.stoa.org/gl/latest/. ((2007-2022)), accessed January 22, 2022

[8] Jaccard: The distribution of the flora of the alpine zone. In: New Phytologist. vol. 11, pp. 37–50 (1912)

[9] Jacobs, J.: Finding words that sound alike. The SOUNDEX algorithm. Byte 7 pp. 473–474 (1982)

[10] Lee, M.L., Yang, L.H., Hsu, W., Yang, X.: Xclust: clustering xml schemas for effective integration. In: Proceedings of the eleventh international conference on Information and knowledge management. pp. 292–299 (2002)

[11] Melzer, S., Peukert, H., Wang, H., Thiemann, S.: Model-based Development of a Federated Database Infrastructure to support the Usability of Cross-Domain Information Systems. In: IEEE International Systems Conference (SysCon 2022), Montreal, Canada. IEEE (2022)

[12] Melzer, S., Schiff, S., Weise, F., Harter, K., Möller, R.: Databasing on demand for research data repositories explained with a large epidoc dataset. CENTERIS (2022)

[13] Melzer, S., Thiemann, S., Möller, R.: Modeling and simulating federated databases for early validation of federated searches using the broker-based sysml toolbox. In: IEEE International Systems Conference, SysCon 2021, Vancouver, BC, Canada, April 15 - May 15, 2021. pp. 1–6. IEEE (2021)

[14] Miller, F.P., Vandome, A.F., McBrewster, J.: Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau-Levenshtein Distance, Spell Checker, Hamming Distance. Alpha Press (2009)

[15] Odell, M.K., Russell, R.: Patent numbers 1261167 (1918) and 1435663 (1922). Washington, DC: US Patent Office (1918)

[16] OpenAI: ChatGPT (Vers. 3.5). https://openai.com (2021)

[17] Pergantis, M., Varlamis, I., Giannakoulopoulos, A.: User Evaluation and Metrics Analysis of a Prototype Web-Based Federated Search Engine for Art and Cultural Heritage. Information **13**(6), 285 (Jun 2022)

[18] Rahm, E., Do, H.H., Massmann, S.: Matching large xml schemas. ACM SIGMOD Record **33**(4), 26–31 (2004)

[19] Schiff, S., Melzer, S., Wilden, E., Möller, R.: TEI-Based Interactive Critical Editions. In: Uchida, S., Barney, E., Eglin, V. (eds.) Document Analysis Systems. pp. 230–244. Springer International Publishing, Cham (2022)

[20] Shokouhi, M., Baillie, M., Azzopardi, L.: Updating Collection Representations for Federated Search. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 511–518. SIGIR '07, Association for Computing Machinery, New York, NY, USA (2007)

[21] Shokouhi, M., Si, L.: Federated Search. Foundations and Trends® in Information Retrieval **5**(1), 1–102 (2011)

[22] Text Encoding Initiative: P5: Guidelines for Electronic Text Encoding and Interchange, Version 4.0.0. https://tei-c.org/Vault/P5/4.0.0/doc/tei-p5-doc/en/html/ (2020), accessed 29 June 2022

[23] Universität Hamburg: Epigraphische Datenbank zum antiken Kleinasien (2013-2016), https://www.epigraphik.uni-hamburg.de/content/index.xml

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is All you Need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)