

Hajra, Arben; Radevski, Vladimir; Tochtermann, Klaus

Conference Paper — Accepted Manuscript (Postprint)

Author Profile Enrichment for Cross-linking Digital Libraries

Suggested Citation: Hajra, Arben; Radevski, Vladimir; Tochtermann, Klaus (2015) : Author Profile Enrichment for Cross-linking Digital Libraries, In: Kapidakis, Sarantos Mazurek, Cezary Werla, Marcin (Ed.): Research and Advanced Technology for Digital Libraries. 19th International Conference on Theory and Practice of Digital Libraries, TPD 2015, Poznan, Poland, September 14-18, 2015, Proceedings, ISBN 978-3-319-24592-8, Springer International Publishing, Cham, Switzerland, pp. 124-136,
https://doi.org/10.1007/978-3-319-24592-8_10
This Version is available at:
<http://hdl.handle.net/11108/231>

Kontakt/Contact

ZBW – Leibniz-Informationszentrum Wirtschaft/Leibniz Information Centre for Economics
Düsternbrooker Weg 120
24105 Kiel (Germany)
E-Mail: info@zbw.eu
<http://zbw.eu/de/ueber-uns/profil/veroeffentlichungen-zbw/>

Standard-Nutzungsbedingungen:

Dieses Dokument darf zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden. Sie dürfen dieses Dokument nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Sofern für das Dokument eine Open-Content-Lizenz verwendet wurde, so gelten abweichend von diesen Nutzungsbedingungen die in der Lizenz gewährten Nutzungsrechte.

Terms of use:

This document may be saved and copied for your personal and scholarly purposes. You are not to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. If the document is made available under a Creative Commons Licence you may exercise further usage rights as specified in the licence.

Author Profile Enrichment for Cross-linking Digital Libraries

Arben Hajra¹, Vladimir Radevski¹, Klaus Tochtermann²

¹ South East European University (SEEU), Tetovo, R. of Macedonia
{a.hajra, v.radevski}@seeu.edu.mk

² Leibniz Information Centre for Economics (ZBW), Kiel, Germany
{k.tochtermann}@zbw.edu

Abstract. This work aims at enriching author profiles with additional information to better support search and retrieval of publications across different digital libraries. To achieve this objective we exploit concepts for cross-linking data to identify correlations between one author and other authors, publications or other related information. We will introduce a profile enrichment approach which adds additional information (e.g. biographic information) from different sources to existing author profiles. Within this context, the linked open data repository DBpedia serves a valuable source for our profile enrichment approach. Still, one of several challenges in this context is the identification of the same author in different sources. To address this challenge we will exploit VIAF (virtual authority file) for author identification. Technically we apply data mining and clustering techniques to uniquely identify authors.

Keywords: Digital libraries, VIAF, author disambiguation, data mining, profile enrichment, linked open data.

1 Introduction

It is widely accepted that Digital Libraries (DL) play an important role in modern scholarly communication. Typically DLs hold domain specific information (e.g. economics) which makes it difficult to search across different domains. For example, would a scholar need literature from economics and agriculture he or she would have to access two different DLs. To overcome this limitation, we will explore options for achieving interoperability by cross-linking authors and/or publications from different DLs with one another.

The main aim of our work is enriching the content of a DL with additional information from other DLs especially regarding information which is somehow related to the authors. Our primary objective is as follows: Assume we have found publications and bibliographic information from an author in one DL, we want to harvest other DLs for correlations to other publications of the same authors, of his or her co-authors and for additional bibliographic information of the initial author.

Our approach suggests creating an author profile, based on the information we have collected from one DL. This profile will continuously be enriched with addition-

al information found in other DLs. To enrich the search results from one DL with additional results from other DLs we apply author name disambiguation, author identification and false authorship prevention.

To uniquely identify authors and to create correlations between them, we consider bibliographic repositories offered by several libraries and institutions. Very promising is data which are presented in the form of Linked Open Data (LOD), as part of the LOD cloud [1], [3], [4]. As a test case, we will leverage the following repositories: German National Library - DNB, Library of Congress - LC, National Library of France - BNF, National Library of Sweden - KB / LIBRIS.

Finally, we put the Virtual International Authority File - VIAF¹ in the center of our work and utilize it as a “bridge” to those DLs we want to cross-link with each other.

The reminder of this paper is structured as follows. In Section 2 we put our work in context with related work. As a contribution to theory and practice of digital libraries, Section 3 and 4 introduce formally our concepts for profile enrichment, i.e. we present how we collect information for author profiles, how we model them and how we correlate them with one another using VIAF as a bridge. Section 5 shows the practical implementation of our work. It is followed by Section 6 which highlights the most important evaluation results. The paper closes with an outlook on our future work.

2 Related Work

In general, author disambiguation includes two main steps, measuring the similarity and clustering similar records [7]. The main challenge is the identification of whether two authors in the same or different DLs have the same identity or not. The most explored strategies consider the string processing approach which measures the similarity of authors’ names [8], [9]. The comparisons are one-to-many and many-to-many, by applying iterative methods [10]. The explored disambiguation process is generally divided using the following approaches: supervised with heuristic similarity functions, unsupervised and hybrid [6], [11]. In our approach, the similarity measurement is not only based on the author’s names. We also consider the semantic distance between publication titles, co-authors correlations and co-authors publications. As a result, we suggest a completely automated unsupervised clustering technique.

The most explored strategies in the center of the process apply similarity measurements by employing data mining algorithms for text based distances. The data are represented as vector space model where the distance between vectors represents the similarity. Such algorithms include the Cosine Similarity (CS) with TF-IDF, Jaccard Similarity, Jaro Winkler, and Levenshtein algorithms [7], [9], [12], [13], [14].

In almost all these strategies, the author disambiguation process is primarily based on relationships among co-authors and similarity of publications, by discovering other relationships in other DLs [15]. The approach presented in [12], gathers information from citations and submits queries to a Web search engine with the aim to find relevant information about authors. That is, the possibility of user feedback is emphasized

¹ <http://viaf.org/>

on ambiguous references across iterations in which the feedback in combination with the hybrid supervised process is applied for assigning references to authors [13].

Additionally, there are several efforts for generating authority profiles for uniquely identifying resources and researchers. We emphasize: ORCID, VIAF, VIVO, RESEARCHERID and OPENID as most appropriate approaches facilitating the disambiguation of authors and which are used as a “bridge” for retrieving accurate information from different repositories

ORCID - Open Researcher and Contributor Identifier create and maintain a registry of unique researcher identifiers and a method of linking research activities. Main contributors are several publishing houses, scientific communities and universities. It has available APIs under an open source license [18].

VIAF - Virtual International Authority File hosted by OCLC (Online Computer Library Center, Inc.) is a service that virtually integrates multiple authority files from several national libraries into a single OCLC name authority service. VIAF began as a common project with the LC, DNB BNF and OCLC [19].

VIVO - enables the discovery of researchers across institutions. It is an open source semantic web application where through it, institutions such as Cornell, Harvard, and Indiana University, manage and publish information about researchers and their activities [20].

RESEARCHID – to identify potential collaborators and avoid author misidentification, each member is assigned a unique identifier to enable researchers to manage their publication lists. The ResearcherID information integrates with the Web of Science of Thomson Reuters Company [21].

OPENID – is a foundation that promotes OpenID technologies. OpenID Foundation members include leading companies and individuals in the digital identity industry such as Google, Microsoft and Yahoo [22]. Even though this currently has no direct application in the scholarly communication, there is a promising potential.

In our work, we consider VIAF with the highest usage relevance. The main idea of VIAF is to link authority files from several national libraries into a “super” virtual authority record, i.e., cluster. Currently, the most known national libraries maintain their own authority files, which brings a distinctive way of preserving them [19]. The VIAF API can be used by anyone without the need of authentication. In addition, there is also the option of VIAF LOD repositories. However, VIAF strongly recommends the usage of API because of the frequency of updates of the VIAF content.

VIAF links disparate names for the same person by integrating authority files from 35 national libraries from 30 countries into a particular cluster. Each cluster is assigned with a unique number, a VIAF ID. However, there are cases when the VIAF clustering algorithm shows deficiencies, such as: several clusters for the same person, different people into the same cluster, incorrect bibliographic data and clusters with poor content [23]. Based on the results from [17] in a search of 283,114 names, 59% were not ambiguous, meaning that only one heading was found, 26% matched two headings, 10% matched three headings, 3% matched four and 2% more than four.

3 Basic Principles for Profile Enrichment

Our primary goal is enriching the content of a Digital Library with content from other repositories by cross-linking information related to authors. Our research is based on the EconStor² repository, the leading German Open Access repository for economics which is maintained by ZBW. EconStor content has also been published in the LOD.

For each EconStor author, we harvest several other repositories for correlations with other authors, publications or other relevant information about the initial author. As a result, we create a wider author profile enriched with additional information. This profile serves two purposes, to enrich the search result and to solve author ambiguities by global identification of the same author written in different ways or same name referring to different authors.

The process of correct author identification in different repositories is related to the challenge of author's name ambiguity, when determining if two or more references correspond to the same person [2], [5], [6]. For example, an author can be represented with different spellings in several bibliographic repositories or different authors can share the same name, which increases the complexity to the data cross-linking process.

Considering the fact that EconStor content is represented as RDF statements, i.e., linked open data, we extend our interest to other bibliographic repositories in the LOD cloud. Still, the author name ambiguity remains to be the major obstacle for direct information retrieval about a given author from these repositories.

As an example, we would like to find as much information as possible about an EconStor author by harvesting other repositories. We often encounter cases in which the same author is presented with different spelling variations, such as: **Adam Smith; Smith, Adam; A. Smith-; Smith, Adam, 1723-1790; Смит, Адам, 1723; Smith. A.; Smith, Adam T. ; and Smith, Adam, 1930.** In addition, there could be different authors all with the name **Adam Smith**. In principles, a similar problem concerns the metadata about titles of publications which can vary across different repositories.

3.1 EconStor Metadata

The process for data cross-linking is based and initiated from the *metadata* that are used to describe the authors and publications in EconStor. The most basic metadata for describing an author are Name and Surname. An author $\mathbf{a}(\mathbf{a}_{\text{name}}, \mathbf{a}_{\text{surname}})$ is represented by the vector $\mathbf{a} = (t_1, t_2)$. Given this, the set of publications where \mathbf{a} is author is represented as $P_a = \{p_1^a, p_2^a, p_3^a, \dots, p_k^a\}$. Consequently, every certain publication will be composed by the set of terms (strings) found in the title, such: $p_i^a = \{t_1^{pi}, t_2^{pi}, t_3^{pi}, \dots, t_m^{pi}\}$.

Accordingly, for each publication from P_a , other authors are considered to be co-authors of \mathbf{a} . The union of authors from all P_a publications, will represent the set of co-authors, which are denoted as $A_a = \{a_1^a, a_2^a, a_3^a, \dots, a_n^a\}$.

² <http://www.econstor.eu/>

The set of co-authors' publications is of particular importance for determining the co-authorships at the initial repository. With \bar{P}_a we will represent the set of publications of co-authors of \mathbf{a} , where $\bar{P}_a = \{\bar{p}_1^{a1}, \dots, \bar{p}_k^{a1}, \bar{p}_1^{a2}, \dots, \bar{p}_k^{a2}, \bar{p}_1^{a3}, \dots, \bar{p}_k^{an}\}$. Thus, $\bar{P}_a = \{\bar{p}_j^{ai}; i = 1, n; j = 1, k\}$.

Table 1a represents the set of these metadata. A detailed picture of the relationships is shown in Figure 1, where can be seen that p_1^a and p_2^a have a common author.

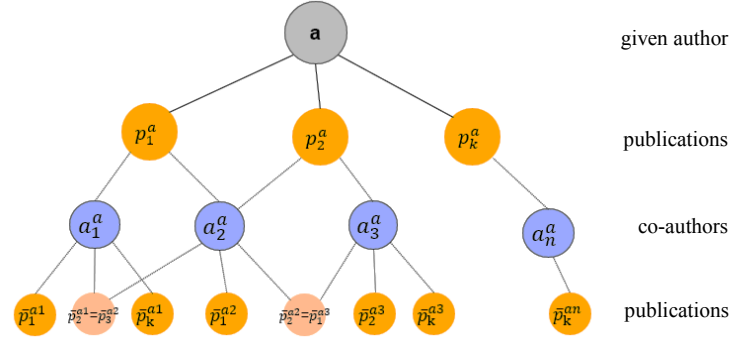


Fig. 1. Relationship among authors, co-authors, publications and co-authors publications for a given author \mathbf{a}

3.2 VIAF Metadata

VIAF clusters are considered as the target repository in which metadata are analyzed. The similarity measurement will be performed between the metadata from the VIAF clusters and the metadata from our repository. For an input author in VIAF the output is delivered by a set of clusters for that author, denoted as c_j , where $j=1, k$. Inside each of these VIAF clusters different forms of authors' name presentations can be found for a particular author, obtained from the native libraries. In this paper, the set of variations is denoted $A_{c_j} = \{a_1^{c_j}, a_2^{c_j}, a_3^{c_j}, \dots, a_l^{c_j}\}$, where each $a_1^{c_j} = (t_1, t_2)$, similarly as in the initial repository. Except this information, in any cluster c_j , a possible list of publications can be found in addition to the list of co-authors assigned to that author. The set of publications found in a particular cluster is notated with $P_{c_j} = \{p_1^{c_j}, p_2^{c_j}, p_3^{c_j}, \dots, p_k^{c_j}\}$, while the set of co-authors inside a cluster will be $\hat{A}_{c_j} = \{\hat{a}_1^{c_j}, \hat{a}_2^{c_j}, \hat{a}_3^{c_j}, \dots, \hat{a}_n^{c_j}\}$.

Besides these data, the set of publications retrieved directly from the libraries or institutions that are contributing in that cluster can be of a particular importance. These publications can be retrieved by referring the identification number of each library for that cluster. Thus, the set of publications extracted from all the sources like this, are presented with the set $\check{P}_{c_j} = \{\check{p}_1^{c_j}, \check{p}_2^{c_j}, \check{p}_3^{c_j}, \dots, \check{p}_k^{c_j}\}$. Table 1b represents the set of metadata from a particular VIAF cluster that we are considering.

Table 1a. Notation table - metadata from the initial repository

$\mathbf{a}, \mathbf{a} = (t_1, t_2)$	the author to be disambiguated
$P_a = \{p_1^a, p_2^a, p_3^a, \dots, p_k^a\}$	publications of author \mathbf{a}
$p_i^a = \{t_1^{pi}, t_2^{pi}, t_3^{pi}, \dots, t_m^{pi}\}$	title's terms from the publication
$\Lambda_a = \{a_1^a, a_2^a, a_3^a, \dots, a_{n_2}^a\}$	co-authors of the author \mathbf{a}
$\bar{P}_a = \{\bar{p}_1^{a1}, \dots, \bar{p}_k^{a1}, \dots, \bar{p}_1^{an}, \dots, \bar{p}_k^{an}\}$	publications of co-authors of \mathbf{a}

Table 1b. Notation table - metadata from a VIAF cluster

c_j	clusters to be checked at VIAF
$\Lambda_{c_j} = \{a_1^{c_j}, a_2^{c_j}, a_3^{c_j}, \dots, a_l^{c_j}\}$	author's names variations in a VIAF cluster c_j , $j=1, k$
$P_{c_j} = \{p_1^{c_j}, p_2^{c_j}, p_3^{c_j}, \dots, p_k^{c_j}\}$	publications in a VIAF cluster c_j
$\hat{\Lambda}_{c_j} = \{\hat{a}_1^{c_j}, \hat{a}_2^{c_j}, \hat{a}_3^{c_j}, \dots, \hat{a}_n^{c_j}\}$	co-authors in a VIAF cluster c_j
$\bar{P}_{c_j} = \{\bar{p}_1^{c_j}, \bar{p}_2^{c_j}, \bar{p}_3^{c_j}, \dots, \bar{p}_k^{c_j}\}$	publications from other sources in the VIAF cluster

4 Application of the Profile Enrichment

In our work we consider VIAF as a "bridge" to cross-link different bibliographic repositories. It is a challenge to detect accurately a particular author from a repository, i.e., EconStor, and to connect this author with the corresponding author in VIAF. Achieving the right identification will facilitate the process of retrieving information from other repositories, especially from libraries that contribute to VIAF records, such as, DNB, LC, BNF and LIBRIS. We also consider other publications from a given author, correlations with co-authors, biographical data, publishers, etc.

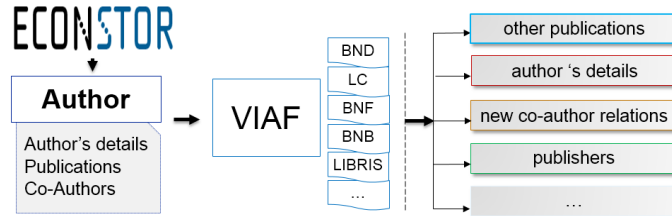


Fig. 2. The overview for enriching process with additional information about authors

4.1 Identifying Authors in VIAF

In section 3.2 we highlighted that a search in VIAF results in several records which match the name of an author. In a second step, we assess the accuracy for each retrieved record.

For this purpose, we implement data mining techniques, by adopting different vector space algorithms. With highest priority, we use the Cosine Similarity (CS) in combination with TF-IDF for the distance between publications, while we apply Levenshtein distance and Jaro distances for similarity author names. The algorithm we propose follows ideas from the process of name deduplication and address information [24].

We start by defining the metadata for the publications in our native repository. These metadata are described in detail in Section 3. In the very beginning, the process starts by using the VIAF API for identifying a particular author. Each retrieved cluster is analyzed in iterative fashion according to these steps:

i. Similarity among author's name with the alternatives within a cluster.

In cases when at least one full match is found, a particular weight is assigned to the variable, denoted as w_{ac} . In detail, the similarity check is done only in the context of the authors name and surname as terms in a vector, i.e. $\mathbf{a} = (t_1, t_2)$ and $a_i^{cj} = (t_1, t_2)$. Thus, iteratively for each name alternative a_i^{cj} within a cluster, similarity measurement is calculated with the author \mathbf{a} .

$$w_{ac} = \text{sim}(a, a_i^{cj}), i = 1, n; j = 1, k; \quad (1)$$

The similarity among names in this step is calculated with CS and TF-IDF where only the perfect match among names is considered. We take this simplified approach to avoid any unreliable results that could be infiltrated when otherwise.

ii. Similarity between publications that an author has in our repository with the publications found in the VIAF cluster.

With \mathbf{P}_a is assigned the set of all publications that this author has in our repository, while with \mathbf{P}_{cj} the set of publications found in a particular cluster. Each publication from our repository is compared with each publication found in the cluster. The similarity between publications can be measured based on Cosine Similarity with TF-IDF, where each publication is presented as an array of strings, i.e., terms that consist of the title of the publication. The outcome of CS is bounded between 0 and 1, where 1 represents a complete match. Thus, a publication $p_e^a \in P_a$, $p_e^a = \{t_1^{pi}, t_2^{pi}, t_3^{pi}, \dots, t_k^{pi}\}$ and $p_f^{cj} \in P_{cj}$, $p_f^{cj} = \{t_1^{cj}, t_2^{cj}, t_3^{cj}, \dots, t_m^{cj}\}$ we have:

$$w_{pc} = \text{sim}(p_e^a, p_f^{cj}), e = 1, k; f = 1, m; k, m \geq 3; \quad (2)$$

In this case for each comparison a specific weight w_{pc} is assigned. Its value is determined if the similarity among the compared titles is above the defined threshold, which is 0.6 for publications that have more than three terms in the title. This value is set based on our preliminary analysis, which showed that lower thresholds and less than three terms in the title, resulted in inaccurate matching.

Before performing the similarity algorithm, the cleaning and formatting of the data is conducted, such as: removing punctuation, eliminating “stopwords”, lowercase and encoding the data to Unicode character encoding (UTF-8).

iii. Comparing the list of co-authors for an author with co-authors found in the cluster.

Let us consider $A_a = \{a_1^a, a_2^a, a_3^a, \dots, a_n^a\}$ the set of co-authors with whom the author \mathbf{a} has at least one common publication, while $\hat{A}_{cj} = \{\hat{a}_1^{cj}, \hat{a}_2^{cj}, \hat{a}_3^{cj}, \dots, \hat{a}_n^{cj}\}$ is the set of co-authors in a particular VIAF cluster c_j . In this case, as it is explained in (ii), each co-author from A_a is compared with each co-author from \hat{A}_{cj} .

$$w_{\hat{c}} = \text{sim}(a_e^a, \hat{a}_f^{cj}), e = 1, k; f = 1, m; \quad (3)$$

At least one match, $A_a \cap \hat{A}_{cj} \neq \emptyset$, can be a significant proof that our repository and the cluster have a common co-author. In that case variable $w\hat{a}_c$ will get a weight for each iteration in this step. Having more than one match increases the evidence that it is the required cluster. A more suitable similarity metric for names is applied based on the Jaro-Winkler similarity metric. In this case the similarity is calculated according to the characters. The threshold for names calculated by CS remains 1.0, while for Jaro-Winkler it will be above 0.9.

iv. Checking the list of publications directly from the sources (libraries) that belong to the cluster.

The set of publications retrieved from the libraries that belong to the cluster c_j , is denoted with \check{P}_{cj} . For example, if DBN has its records in that cluster, we are measuring the similarity between them and publications from our repository, $p^a \in P_a$ with $\check{p}^{cj} \in \check{P}_{cj}$. For each check, a particular weight is assigned to the variable $w\check{p}_c$, absolutely in the same manner as in the step (ii).

$$w\check{p}_c = \text{sim}(p_e^a, \check{p}_f^{cj}), e = 1, k; f = 1, m; k, m \geq 3; \quad (4)$$

4.2 Determining the Matching Degree

The key factors for determining the matching degree between an author from our repository with a particular VIAF cluster, are precisely the components presented above. At each of these components, under (i), (ii), (iii) and (iv) the weight is calculated iteratively with equations (1), (2), (3) and (4). The overall “weight” is calculated in accumulative way such as $\text{sum}(w_{ac}, w_{pc}, w_{\hat{a}_c}, w_{\check{p}_c})$, by respecting the threshold.

5 Sample Implementation for Profile Enrichment

In this section we describe the prototype used for the evaluation of the developed algorithms. This prototype automatically checks VIAF for a particular author and automatically determines the appropriate clusters according to the principles presented in the previous sections. For each cluster found, the VIAF ID is taken and assigned to the corresponding author in the initial repository (EconStor in our case). As a result, an author’s profile is enriched with additional information found in the cluster.

For the implementation we use EconStor and an RDF dump file of Econstor. EasyRdf PHP library and rdf4j Sesame are applied for processing and storing the RDF data. The current version of dump file contains 1.635.599 RDF statements, 36.490 publications and 27.580 authors.

To give an example, we select a particular author, i.e. “Kubler, Felix”, in EconStor. As a result a list of all publications, co-authors and co-author’s publications from our repository will be created and returned to the user of our prototype. Considering this author, the prototype found six clusters in VIAF, of which the third one is depicted in Figure 3. In this cluster, similarities are found related to the author’s name, publications, co-authors and publications from the libraries that belong to it. From the list of six publications, the prototype has highlighted three with 100% match to the EconStor publication. Also, four co-authors of “Kubler, Felix” with 100% match were found.

3. **Kubler, Félix** - (<http://viaf.org/viaf/57719599>)

Publications:

1. Borrowing costs and the demand for equity over the life cycle - 100 %
2. Collateralized borrowing and life-cycle portfolio choice - 100 %
3. Computational aspects of general equilibrium theory : refutable theories of value - 31.45 %
4. Computing stochastic dynamic economic models with a large number of state variables: a description and application of a Smolyak-collocation method - 35.86 %
5. The robustness of the CAPM-A computational approach - 50.25 %
6. Social security and risk sharing - 100 %

Total similarity between titles: 68.15

Co-Authors:

1. Krueger, Dirk 100 %
2. Willen, Paul 100 %
3. Malin, Benjamin. 0 %
4. Herings, P. Jean-Jacques (Peter Jan Jacob), 1969- 0 %
5. Gottardi, Piero 100 %
6. KUBLER, Felix 0 %
7. Chiappori, P. -A. 0 %
8. Brown, Donald J. 0 %
9. SCHMEDDERS, Karl 100 %
10. KUBLER, F. 0 %

Sources:

1. NTA|183780418 - 183780418
2. DNB|130434612 - <http://d-nb.info/gnd/130434612>
3. SUDOC|129614262 - 129614262
4. LC|no2003007063 - no2003007063
5. ISNI|0000000044322045 - 0000000044322045

Living: 0-0

Other links:

- * [GNDB links](#) - and a list of possible [Publications from GNDB](#)
- Social security and risk sharing - 100 %
- <http://www.idref.fr/129614262/id>




Fig. 3. The case in which the prototype found and evaluated as correct match an EconStor author with a VIAF cluster

Additionally, there are in total five libraries (“Sources” in Fig. 3) or institutions which contain this cluster, thus a possible exploration in these resources would endorse the match. For example, in the German National Library, a publication is found with 100% similarity (see “Other links” in Fig. 3). However this result is excluded from the calculation because the same publication appears in the cluster’s publications, $p^a = p^{cj}$ (publication 6 in Fig. 3). Overall, all these elements provide evidence that this cluster is correct for the author “Kubler, Felix”.

For a performed search the number of retrieved results can vary from zero to some hundreds. The above example had only six clusters, with only one correct cluster. However, there are several cases in which for one author the number of correct clusters can be zero, one or more than one cluster that really represents him. In case that at least one cluster is found, the VIAF ID is saved in our local database, for each author.

6 Evaluation

We have randomly analyzed 991 authors from EconStor to VIAF and generated the evaluation metrics of recall, precision and F1 score. In our case, precision represents the fraction of the clusters that are retrieved as correct match. In fact it is the fraction among the truly correct clusters (true positive) with all clusters that the system has

retrieved as correct, including clusters that are retrieved as correct but are not (false negative).

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} = \frac{|\{\text{truly correct clusters}\}|}{|\{\text{all retrieved clusters as correct}\}|}$$

The recall represents the fraction between the truly correct clusters with all correct clusters, including the clusters that are correct but the system has not identified them as such (false negative).

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} = \frac{|\{\text{truly correct clusters}\}|}{|\{\text{all correct clusters}\}|}$$

Based on the manually checked evaluations the system gives an overall precision of 98.1% and the recall of 95.9%. Thus, the efficiency of our system is measured with 0.970 as F1 score.

The results in Table 2 represent only the clusters that are marked as positive and the prototype has marked them as correct clusters. However, there are cases in which for an author only one or more than one clusters are retrieved as correct match.

Table 2. The number of found VIAF clusters for EconStor authors.

Number of checked authors from EconStor	Number of truly found clusters in VIAF	%	Precision	Recall	F1
for 598	1	60.3%	0.988	0.957	0.972
for 125	2	12.6%	0.957	0.972	0.964
for 18	3	1.8%	0.952	0.976	0.964
for 9	> 3	0.9%	0.951	0.978	0.964
for 241	0	24.3%	/	/	/

Each of these found clusters are manually evaluated for accuracy of matches. Based on these evaluations, very satisfactory results are generated. In the cases when an author is matched with only one VIAF cluster, we gain 98.8% precision, 95.7% recall and F1 score of 0.972. Thus the possibility for it to be the correct cluster is almost absolute. In the cases when two clusters are retrieved as correct match for one author, the precision is 95.7% and 97.2% recall, with F1 score of 0.964.

For each checked author from our repository, the corresponding VIAF ID is stored locally. Grouping authors like this can be a huge benefit for clustering them inside a local repository and for creating a local authority profile. Beyond this, the found VIAF ID offers a permanent link to that cluster in VIAF. This avoids to repeat the process of identification again. With the right VIAF ID, all the relevant information found in the cluster are instantly retrieved, such as new publications and new co-authorship correlations. Figure 3 shows an example of this.

In addition, each cluster keeps in it the identification number of libraries or institutions that are contributing with content. We are considering these IDs as valuable information for extending the enrichment of an author profile. Therefore, by having that id, such as *13043612* for DNB, *129614262* for SUDOC, we can refer directly to these repositories to search this author. This can be done by different Web Services and APIs which these libraries offer, or by querying the LOD repositories. Most know libraries including DNB, LC, BNE, BNB, BNF, and LIBRIS offer their data or

metadata as LOD in LOD cloud. Consequently, by performing a SPARQL query in these repositories, direct information retrieval is possible.

In several cases, a particular VIAF cluster offers alignment to DBpedia for the corresponding author. We consider this as a possibility to extend an author profile with several other information. The prototype automatically realizes a SPARQL query in DBpedia and retrieves information such as: a short bio, an author picture, a link to Wikipedia page and a downloadable list of works. Figure 4 depicts details from the output of this process.

* http://dbpedia.org/resource/Lars_Peter_Hansen_DBPEDIA


Abstract:	Lars Peter Hansen (born October 26, 1952) is the David Rockefeller Distinguished Service Professor of economics at the University of Chicago. Best known for his work on the Generalized Method of Moments, he is also a distinguished macroeconomist, focusing on the linkages between the financial and real sectors of the economy. In 2013, it was announced that he would be awarded the Nobel Memorial Prize in Economics, jointly with Robert J. Shiller and Eugene Fama.
Birth Date:	1952-10-26+02:00
Picture:	
Wikipedia:	WIKIPEDIA Link
Other Links:	Link: http://usu.edu/ust/index.cfm?article=51098 Link: http://www.mfmgroupp.org/lars-peter-hansen-8.html Link: http://cowles.econ.yale.edu/conferences/koopmans/tck08 Link: http://home.uchicago.edu/~lhansen/ Link: http://home.uchicago.edu/~lhansen/LarsHansenInterviewJBES.pdf

Fig. 4. Finding and extracting author's information from DBpedia

7 Conclusion and Future Work

Relying on the initial idea of creating enriched author profiles in a digital library by extracting data from several other repositories, the process of author disambiguation is inevitable. We referred to VIAF for avoiding ambiguity and uniquely identifying each author from our repository. Note, that our algorithm is not limited to EconStor only; it should work for any repository given that the following input data are provided: author name, list of publications, co-author names and their publications.

Using our promising results, author profiles as part of a digital library can be enriched by useful information such as new publications which are not part of the initial repository, new co-authorship correlations, publications of co-authors, possibility to cluster authors in the initial repository, biographic information, and DBpedia content.

As future work, improvements in the process of similarity measurements will be performed. This will be done by incorporating and combining several metadata elements and by performing other analyses for similarity calculations. Such analyses will impact the process of threshold calculations and consequently improve the determinations of a cluster's accuracy.

References

1. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web. In Proceedings of the 17th international conference on World Wide Web pp. 1265-1266. ACM (2008)

2. Elmagarmid, A. K., Ipeirotis, P. G., Verykios, V. S.: Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1), 1-16 (2007)
3. Hajra, A., Latif, A., Tochtermann, K.: Retrieving and ranking scientific publications from linked open data repositories. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business (I-Know)*, p. 29, ACM (2014)
4. Latif, A., Borst, T., Tochtermann, K.: Exposing Data from an Open Access Repository for Economics as Linked Data. *D-Lib Magazine*, 20(9/10), (2014).
5. Laender, A. H., et al.: Keeping a digital library clean: new solutions to old problems. In *Proceedings of the eighth ACM symposium on Document engineering*, Sao Paulo, Brazil pp. 257-262. ACM (2008)
6. Santana, A. F., Goncalves, M. A., Laender, A. H., Ferreira, A.: Combining domain-specific heuristics for author name disambiguation. In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries*, pp. 173-182. IEEE (2014)
7. Chin, W. S., et al.: Effective string processing and matching for author disambiguation. *The Journal of Machine Learning Research*, 15(1), 3037-3064 (2014)
8. Torvik, V. I., Smalheiser, N. R.: Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3), 11. (2009)
9. Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., Fienberg, S.: Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5), 16-23. (2003)
10. Bhattacharya, I., Getoor, L.: Iterative record linkage for cleaning and integration. In *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery* (pp. 11-18). ACM (2004)
11. Tang, J., Fong, A. C. M., Wang, B., Zhang, J.: A unified probabilistic framework for name disambiguation in digital library. *Knowledge and Data Engineering, IEEE Transactions on*, 24(6), 975-987. (2012)
12. Pereira, D. A., et al.: Using web information for author name disambiguation. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* pp.49-58. ACM (2009)
13. Godoi, T. A., et al.: A relevance feedback approach for the author name disambiguation problem. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* pp. 209-218. ACM (2013)
14. Fan, X., Wang, J., Pu, X., Zhou, L., Lv, B.: On graph-based name disambiguation. *Journal of Data and Information Quality (JDIQ)*, 2(2), 10. (2011)
15. De Nies, T., et al.: Towards named-entity-based similarity measures: Challenges and opportunities. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval* pp. 9-11. ACM (2014)
16. Mazov, N. A., Gureev, V. N.: The role of unique identifiers in bibliographic information systems. *Scientific and Technical Information Processing*, 41(3), 206-210. (2014)
17. Freire, N., et al.: Author Consolidation across European National Bibliographies and Academic Digital Repositories. In *Proceedings of the 11th International Conference on Current Research Information System*. (2012)
18. What is ORCID?, <http://orcid.org/content/about-orcid>
19. Virtual International Authority File, <http://www.oclc.org/viaf/en.html>
20. What is VIVO?, <http://www.vivoweb.org/about>
21. What is ResearcherID?, <http://www.researcherid.com/>
22. OpenID Foundation, <http://openid.net/foundation/>
23. DNB-Virtual International Authority File (VIAF), <http://www.dnb.de/viaf>
24. Bilenko, M., Mooney, R. J., Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* pp. 39-48, ACM (2003)