

Hajra, Arben

Doctoral Thesis — Published Version

Content Enrichment of Digital Libraries: Methods, Technologies and Implementations

Suggested Citation: Hajra, Arben (2020) : Content Enrichment of Digital Libraries: Methods, Technologies and Implementations, Christian-Albrechts-Universität, Kiel,
<https://nbn-resolving.de/urn:nbn:de:gbv:8:3-2021-00210-4>

This Version is available at:
<http://hdl.handle.net/11108/491>

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.

Content Enrichment of Digital Libraries: Methods, Technologies and Implementations

Arben Hajra

Dissertation

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften

Dr.-Ing.

der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel

Kiel, Deutschland, 2020

1. Gutachter: Prof. Dr. Klaus Tochtermann

2. Gutachter: Prof. Dr. Vladimir Radevski

Datum der mündlichen Prüfung: 26. Oktober 2020

Zusammenfassung

Interoperabilität zwischen Bibliotheken ist seit Jahrhunderten eine unzureichend gelöste Schwachstelle, wenn es darum geht, Bibliotheken und ihre Inhalte miteinander zu verknüpfen. Gründe hierfür und für die daraus resultierende Versäulung liegen in unterschiedlichen Bibliothekspraktiken für die Beschreibung und Kuratierung von Metadaten. Diese Unterschiede ergeben sich insbesondere aus der Domäne, der Ressourcenstruktur, den unterschiedlichen Katalogisierungsregeln, der Anwendung unterschiedlicher Metadatenschemata, Ontologien und Vokabulare. Dank der Digitalisierung können Digitale Bibliotheken diese Versäulung von bibliothekarischen Datenquellen bis zu einem gewissen Grad mildern. Gleichzeitig entstehen jedoch weitere Herausforderungen, da mit der Digitalisierung von Bibliotheken, auch Erwartungen der Benutzer*innen an Zugänge zu nicht-bibliothekarischen Datenquellen entstehen.

Parallel zur Etablierung des Konzepts einer „Digitalen Bibliothek“ gab es rasante Weiterentwicklungen in den Bereichen semantischer Technologien, Information Retrieval und künstliche Intelligenz. Im Kontext semantischer Technologien sind das semantische Web, Linked-Data und damit verbundene Technologien für persistente Identifikatoren von besonderer Bedeutung für diese Arbeit. Verfahren des Information Retrieval, wie Vektorraummodell und „Word Embedding“ können genutzt werden, um inhaltliche Ähnlichkeiten abzuschätzen. Im Themenfeld künstliche Intelligenz entwickelte spezielle Methoden für maschinelles Lernen haben einen Reifegrad erreicht, der ihren breiten Einsatz beschleunigt.

Die Anwendung und Kombination von semantischen Technologien, Verfahren des Information Retrieval und des maschinellen Lernens bilden den Ausgangspunkt für diese Dissertation. Im Spannungsfeld dieser drei Themenfelder der Informatik positioniert sich diese Dissertation, die sich als Ziel setzt, anwendungsorientierte Beiträge zur Verbesserung der Interoperabilität zwischen Digitalen Bibliotheken aber auch zwischen Digitalen Bibliotheken und nicht-bibliothekarischen Datenquellen zu leisten.

Die Idee ist es, mit ihrer Hilfe bibliographische Daten, also Inhalte von Bibliotheken, miteinander zu vernetzen und „intelligent“ mit zusätzlichen, insbesondere nicht-bibliothekarischen Informationen anzureichern. Durch die Verknüpfung von Inhalten einer Bibliothek wird es möglich, einen Zugang für Benutzer*innen anzubieten, über den semantisch ähnliche Inhalte unterschiedlicher Digitaler Bibliotheken zugänglich werden. Beispielsweise können hierüber ausgehend von einer bestimmten Publikation eine Liste semantisch ähnlicher Publikationen ggf. aus völlig unterschiedlichen

Themenfeldern und aus verschiedenen digitalen Bibliotheken zugänglich gemacht werden. Darüber hinaus können sich Nutzer*innen ein breiteres Autoren-Profil anzeigen lassen, das mit Informationen wie biographischen Angaben, Namensalternativen, Bildern, Berufsbezeichnung, Instituts-Zugehörigkeiten usw. angereichert ist. Diese Informationen kommen aus unterschiedlichsten und in der Regel nicht-bibliothekarischen Quellen. Um derartige Szenarien Realität werden zu lassen, verfolgt diese Dissertation zwei Ansätze.

Der erste Ansatz befasst sich mit der Vernetzung von Inhalten Digitaler Bibliotheken, um auf Basis zusätzlicher Informationen für eine Publikation semantisch ähnliche Publikationen anzubieten. Dieser Ansatz verwendet publikationsbezogene Metadaten als Grundlage. Die verknüpften Begriffe zwischen verlinkten offenen Datenrepositorien/Thesauri werden als wichtiger Angelpunkt betrachtet, indem Unterbegriffe, Oberbegriffe und verwandten Konzepte über semantische Datenmodelle, wie SKOS, berücksichtigt werden. Methoden des Information Retrieval werden angewandt, um v.a. Publikationen mit hoher semantischer Verwandtschaft zu identifizieren. Zu diesem Zweck werden Ansätze des Vektorraummodells und des „Word Embedding“ eingesetzt und vergleichend analysiert. Die Analysen werden in Digitalen Bibliotheken mit unterschiedlichen thematischen Schwerpunkten (z.B. Wirtschaft und Landwirtschaft) durchgeführt. Durch Techniken des maschinellen Lernens werden hierfür Metadaten angereichert, z.B. mit Synonymen für inhaltliche Schlagwörter, um so Ähnlichkeitsberechnungen weiter zu verbessern. Zur Sicherstellung der Qualität werden die beiden Ansätze mit verschiedenen Metadatensätzen vergleichend analysiert wobei die Beurteilung durch Expert*innen erfolgt. Durch die Verknüpfung verschiedener Methoden des Information Retrieval kann die Qualität der Ergebnisse weiter verbessert werden. Dies trifft insbesondere auch dann zu wenn Benutzerinteraktion Möglichkeiten zur Anpassung der Sucheigenschaften bieten.

Im zweiten Ansatz, den diese Dissertation verfolgt, werden autorenbezogene Daten gesammelt, verbunden mit dem Ziel, ein umfassendes Autorenprofil für eine Digitale Bibliothek zu generieren. Für diesen Zweck kommen sowohl nicht-bibliothekarische Quellen, wie Linked Data-Repositorien (z.B. WIKIDATA) und als auch bibliothekarische Quellen, wie Normdatensysteme, zum Einsatz. Wenn solch unterschiedliche Quellen genutzt werden, wird die Disambiguierung von Autorennamen über die Nutzung bereits vorhandener persistenter Identifikatoren erforderlich. Hierfür

bietet sich ein algorithmischer Ansatz für die Disambiguierung von Autoren an, der Normdaten, wie die des Virtual International Authority File (VIAF) nachnutzt.

Mit Bezug zur Informatik liegt der methodische Wert dieser Dissertation in der Kombination von semantischen Technologien mit Verfahren des Information Retrievals und der künstlichen Intelligenz zur Erhöhung von Interoperabilität zwischen Digitalen Bibliotheken und zwischen Bibliotheken und nicht-bibliothekarischen Quellen. Mit der Positionierung dieser Dissertation als anwendungsorientierter Beitrag zur Verbesserung von Interoperabilität werden zwei wesentliche Beiträge im Kontext Digitaler Bibliotheken geleistet: (1) Die Recherche nach Informationen aus unterschiedlichen Digitalen Bibliotheken kann über einen Zugang ermöglicht werden. (2) Vorhandene Informationen über Autor*innen werden aus unterschiedlichsten Quellen eingesammelt und zu einem Autorenprofil aggregiert.

Abstract

Interoperability between libraries has been for centuries an insufficiently solved limitation when it comes to linking libraries and their contents. The reasons for this and the resulting isolation are the different library practices for describing and curating metadata. These differences arise in particular from the domain, resource structure, different cataloging rules, the use of different metadata schemas, ontologies and vocabularies. Thanks to digitization, digital libraries can to some extent mitigate the isolation of library resources. However, at the same time, further challenges arise as the digitization of libraries also raises user expectations of access to non-library data sources.

Parallel to the establishment of the concept of a "digital library", there have been rapid developments in the fields of semantic technologies, information retrieval and artificial intelligence. In the context of semantic technologies, the semantic web, linked data and related technologies for persistent identifiers are of particular importance for this work. Information retrieval techniques, such as Vector Space Models and Word Embedding, can be used to estimate content-based similarities. Special methods for machine learning developed in the field of artificial intelligence have reached a level of maturity that has accelerated their widespread use.

The application and combination of semantic technologies, information retrieval and machine learning methods form the base for this dissertation, which is positioned in the area between these three fields of computer sciences. The goal of this dissertation is to make application-oriented contributions to improve the interoperability between digital libraries, but also between digital libraries and non-library data sources.

The idea is to use make use of these three fields to crosslink bibliographic data, i.e., library content, and to enrich it "intelligently" with additional, especially non-library, information. By linking the contents of a library, it is possible to offer users access to semantically similar contents of different digital libraries. For instance, a list of semantically similar publications from completely different subject areas and from different digital libraries can be made accessible. In addition, the user is able to see a wider profile about authors, enriched with information such as biographical details, name alternatives, images, job titles, institute affiliations, etc. This information comes from a wide variety of sources, most of which are not library sources. In order to make such scenarios a reality, this dissertation follows two approaches.

The first approach is about crosslinking digital library content in order to offer semantically similar publications based on additional information for a publication. Hence, this approach uses publication-related metadata as a basis. The aligned terms between linked open data repositories/thesauri are considered as an important starting point by considering narrower, broader, and related concepts through semantic data models such as SKOS. Information retrieval methods are applied to identify publications with high semantic similarity. For this purpose, approaches of vector space models and "word embedding" are applied and analyzed comparatively. The analyses are performed in digital libraries with different thematic focuses (e.g. economy and agriculture). Using machine learning techniques, metadata is enriched, e.g. with synonyms for content keywords, in order to further improve similarity calculations. To ensure quality, the proposed approaches will be analyzed comparatively with different metadata sets, which will be assessed by experts. Through the combination of different information retrieval methods, the quality of the results can be further improved. This is especially true when user interactions offer possibilities for adjusting the search properties.

In the second approach, which this dissertation pursues, author-related data are harvested in order to generate a comprehensive author profile for a digital library. For this purpose, non-library sources, such as linked data repositories (e.g. WIKIDATA) and library sources, such as authority data, are used. If such different sources are used, the disambiguation of author names via the use of already existing persistent identifiers becomes necessary. To this end, we offer an algorithmic approach to disambiguate authors, which makes use of authority data such as the Virtual International Authority File (VIAF).

Referring to computer sciences, the methodological value of this dissertation lies in the combination of semantic technologies with methods of information retrieval and artificial intelligence to increase the interoperability between digital libraries and between libraries with non-library sources. By positioning this dissertation as an application-oriented contribution to improve the interoperability, two major contributions are made in the context of digital libraries: (1) The retrieval of information from different Digital Libraries can be made possible via a single access. (2) Existing information about authors is collected from different sources and aggregated into one author profile.

Acknowledgments

*"To get the full value of joy
you must have someone to divide it with."*

MARK TWAIN

I would like to express my gratitude and respect to my supervisor, Prof. Klaus Tochtermann, for his invaluable guidance and immense support in every possible direction. I'm deeply thankful to Prof. Vladimir Radevski that generously gave his time and support from the first day of this journey. My sincere thanks also go to the other two members of my exam committee, Prof. Wilhelm Hasselbring, and Prof. Isabella Peters.

A special acknowledgment goes to my colleagues at the Leibniz Information Centre for Economics (ZBW). There I found a great environment for developing and implementing my research ideas, and a fantastic atmosphere for collaborating and engaging discussions. A special thanks to Atif Latif, Fidan Limani, Tamara Pianos, and many others, for the provided feedback in different situations. Likewise, I would like to thank my colleagues from the South East European University (SEEU), in particular Besnik Selimi and Mexhid Ferati, with whom I have discussed and exchanged many ideas and received valuable feedback to my thesis development.

Finally, I would like to thank my parents, my sister, and my brother, for their unparalleled love. Especially indebted to my wife for her limitless support throughout this journey. I am thankful to and blessed with my amazing daughter and son, Era and Genc; you are a marvelous source of inspiration that sprinkles on every word of this work.

Thank you all.

Arben

Brief Contents

1. Introduction	1
1.1 Motivation	1
1.2 Overview of Approaches and Contributions	6
1.3 Publications	10
1.4 Chapters Structure	12
I Foundations	15
2. Scholarly Communication	17
2.1 Digital Libraries	17
2.2 Interoperability of Digital Library resources	19
2.3 Integrated Authority Files	20
2.3.1 Virtual Authority Files (VIAF)	23
2.3.2 WIKIDATA	24
3. Semantic Web and Linked Open Data	27
3.1 Linked Open Data (LOD)	27
3.2 Selected Repositories	30
3.2.1 EconStor Repository	32
3.2.2 AGRIS Repository	33
3.3 Selected Thesauri	33
3.3.1 STW Thesaurus	34
3.3.2 AGROVOC Thesaurus	34
3.3.3 WordNet Thesaurus	35
3.4 Aligned Concepts Between Repositories/Thesauri	35
4. Recommender Systems and Semantic Similarity	37
4.1 Recommender Systems	37
4.2 Similarity scoring	38
4.2.1 Vector Space Model (VSM)	39
4.2.2 Deep Learning through Word Embedding	40
4.3 Ranking evaluation metrics	42

II	Linking Science	43
5.	Research Design	45
5.1	The aim	46
5.2	Identifying sources for enrichment	47
5.3	Proposed approaches	49
5.4	Publication-centered metadata at initial repository	50
5.4.1	Publication-centered metadata at target repository	55
5.5	Author-centered metadata at initial repository	58
5.5.1	VIAF metadata	59
6.	Linking publications across different LOD repositories	63
6.1	Using aligned concepts	64
6.2	Vector Space Model approach	72
6.2.1	Determining the key terms of a publication	72
6.2.2	Measuring the similarity among publications	76
6.2.3	Experimental setup of VSM approach	76
6.2.4	Limitations of VSM	79
6.3	Word Embedding Approach	80
6.3.1	Training and Building the Model	80
6.3.2	Analyzing the Model	81
6.3.3	Experimental setup of Word Embedding approach	83
6.3.4	Limitations of Word Embedding approach	85
6.4	UI integration and scholar involvement	86
6.4.1	Automated Search	88
6.4.2	Customized Search	89
7.	Crosslinking-through author's disambiguation	95
7.1	Author's name disambiguation	97
7.2	Identifying Authors in VIAF	99
7.2.1	Author's name versus alternatives within a cluster	101
7.2.2	Author's publications versus clusters publications	102
7.2.3	Co-authors versus co-authors in the cluster	103
7.2.4	Author's publications versus publications from sources	103
7.3	Determining the Matching Degree	104
7.4	The experimental setup	104
7.4.1	The prototype examples	105
7.4.2	Storing and evaluating the prototype results	111
7.5	Limitations of this approach	113

III	Evaluation and Results	115
8.	Evaluation of approaches across LOD Repositories	117
8.1	Results and Discussions	117
8.1.1	The results	119
8.1.2	Cumulative Gain measures	124
8.2	Summary	127
9.	Evaluation of author's disambiguation	129
9.1	Evaluation of VIAF approach	130
9.2	The outcome after identification	133
9.2.1	Using the VIAF cluster	133
9.2.2	Using cluster's sources	136
9.2.3	Further data enrichment	137
9.2.4	The overview of outcomes	141
9.3	Summary	146
10.	Related Work	147
10.1	Recommender systems from LOD with data mining	147
10.1.1	Linked Open Data in Recommender Systems	147
10.1.2	Vector Space Model and Word Embedding	149
10.2	Authors' disambiguation and identification	150
10.2.1	Authors disambiguation	151
10.2.2	Linking Authors	152
IV	Conclusion and Future Work	153
11.	Conclusion	155
12.	Future Work	161
V	Appendix	163
A.1	The usage of some definitions in our context	165
A.2	The algorithmic approach to disambiguate authors through VIAF...	168
	Bibliography	171

List of Figures

3.1	Semantic Web architecture in layers	29
3.2	Published and interlinked datasets.	31
5.1	The enrichment of a publication or author with other information.....	47
5.2	Extracting publication's metadata from EconStor.....	52
5.3	A descriptor of an EconStor publication.....	53
5.4	Summary of main descriptors by highlighting AGRIS alignments.	54
5.5	The co-authorship network	58
5.6	A particular cluster in VIAF.....	60
6.1	Enriching a scientific publication with information from LOD repositories. 64	
6.2	Thesauri alignments	66
6.3	Alignments to other repositories for a particular STW descriptor.....	67
6.4	Retrieving scientific publications based on concepts' alignments	68
6.5	DBpedia information for a selected STW concept.....	71
6.6	Adjusting the relevance of the metadata components.....	75
6.7	The combination of metadata components to retrieve other publications.....	78
6.8	The similarity measurement scores with CS and Word2Vec.....	84
6.9	Visual Search Interface.....	88
6.10	Customized search	90
6.11	Terms suggested by the thesaurus	91
6.12	Terms suggested from the machine learning approach	92
6.13	Retrieved results based on the visual search approach.....	93
7.1	The author's enrichment approach	96
7.2	The overview for harvesting author's data	99
7.3	Identifying authors in VIAF	100
7.4	Initiating a search for a particular author.....	106
7.5	A correct match between an EconStor author with a VIAF cluster.....	108
7.6	An incorrect match between an EconStor author with a VIAF cluster	109
7.7	The case when the prototype has depicted as "maybe" a VIAF cluster	110
7.8	The prototype and user evaluation for retrieved VIAF clusters.....	111
8.1	Humanly evaluation of top-ten retrieved publications.	121
8.2	The relevance of the retrieved result based on both approaches.	122
8.3	The relevance of the retrieved result versus titles	123
8.4	The DCG and nDCG scores for VSM and WE approaches.	126

9.1	The evaluation results based on the accuracy of the found clusters.	132
9.2	Name variations for a particular author and living year(s).....	133
9.3	The list of publications inside a VIAF cluster for a particular author	134
9.4	The list of co-authors inside a VIAF cluster for a particular author	135
9.5	The list of sources in a particular VIAF cluster	135
9.6	WIKIDATA as authority linking hub	139
9.7	Authority (persons) identifiers in WIKIDATA	140
9.8	An enriched/extended author profile.....	142
9.9	Co-authors network for a particular author	143
9.10	Word Tag cloud with terms from author's publications	144
9.11	Listing authors working on similar topics	145

List of Tables

2.1	The list of some authors identifiers in WIKIDATA.....	25
3.1	The list of properties by vocabularies at EconStor.	32
3.2	STW mappings.....	36
5.1	The notation table – publication’s metadata from the initial repository	54
5.2	A sample of retrieved metadata for an AGRIS publication.	56
5.3	Notation table – author’s metadata from the initial repository	59
5.4	Notation table – author’s metadata from a VIAF cluster.....	61
6.1	A sample of mapped descriptors to another repository, i.e. AGRIS.....	67
6.2	The list of unigrams based on Google Books Ngrams.....	73
6.3	The top-ten most important terms of a publication.	75
6.4	Top-ten most similar words based on the words “inflation” and “food”	82
7.1	The calculated variables to assess the VIAF cluster accuracy match	101
7.2	EconStor authors with the corresponding found VIAF ID.....	112
8.1	An example of the top-ten retrieved and evaluated publications	120
8.2	Generating DCG ₁₀ score on top-ten retrieved publications.	125
9.1	The number of found VIAF clusters for EconStor authors.....	131

Listings

5.1	Retrieving publication's information from EconStor SPARQL endpoint.....	50
5.2	Retrieving publication's author(s) from EconStor SPARQL endpoint.....	51
5.3	Retrieving publication's metadata from the target repository	55
5.4	Retrieving descriptor's label from AGROVOC	56
5.5	Extracting narrowed, broadened and related concepts	57
6.1	Retrieving publications from the target repository	68
6.2	Getting DBpedia information about an STW concept.	70
9.1	Getting other identifiers from WIKIDATA.....	138
9.2	Economic awards for a particular author based on WIKIDATA.....	138
9.3	Getting author's information from DBpedia.	139

Introduction

$$\begin{aligned} 1.01^{365} &= 37.8 \\ 0.99^{365} &= 0.03 \end{aligned}$$

//

The introduction of the thesis begins by highlighting the motivation and problem statement. Afterward, in Section 1.2 the scientific contributions are presented, followed by the list of publications in Section 1.3. Section 1.4 outlines the structure of the thesis, by introducing each chapter.

1.1 Motivation

Traditionally, libraries provide the basic information infrastructures for scholarly communication. As mentioned in [Borg90] and [KIMc99], since the beginning of the 90s, electronic scholarly communication has captured the imagination of many scholars. The era of digitalization emphasized their role in this process, but at the same time, requirements and expectations of services provided by them increased. Thus, libraries are not considered anymore only as a place for finding a particular piece of information, but a place where the required information would be enriched with various data from different places and domains, and lead us to further insights. We would like also to point out to the fact that there is a growing trend of repositories published as linked open data (LOD), WIKIDATA is just one such example that prominently has served as a hub to gather context information for scientific publications and authors.

1. Introduction

Consequently, rather than navigating into the webspace, i.e., several Digital Libraries (DLs) or non-library sources for interlinking relevant information and getting more comprehensive information, the scholar may use a single interface in a preferred DL for that purpose. Hence, a DL would provide automated services for integrating data from various sources and offer scholars the possibility to adjust, tune and filter the data through various facets for further insights and discoveries. In such a case, DLs have successfully managed to adapt to these challenges by improving the utilization of resources from different perspectives, such as quality of services, system performance and user experience [GaGF10, HFCH12, Xie06]. Even so, there is still an evident gap between the demand and supply of DL services to support scholarly use cases [Than16].

Some of these elements that create that gap are related to information retrieval, i.e., recommending semantically related articles (e.g., publications with similar content) based on a preselected publication or set of concepts. The current practices of recommending related articles to a preselected one are mostly based on text matching and word frequencies rather than word relatedness or semantic similarities. As an example, for the publication titled "*Food prices and political instability*" the list of retrieved publications consists of articles where its terms appear in the indexed metadata (i.e., title, abstract, full text, keywords). However, many more publications that are closely related to this will be completely invisible or ranked far below the top publications. This is because the intersections of the metadata terms usually result in an empty set, and there is an inability to identify relatedness among terms such as *energy*, *water* or *fuel* by considering the word *food*. Chapter 6, respectively sections 6.2 and 6.3, presents scenarios and approaches to tackle this particular challenge.

The task becomes even more complex if we are interested to retrieve semantically similar publications from several DLs that belong to different domains. Typically, specialized scientific DLs hold domain-specific information such as economics, social sciences, computer sciences, or agronomics, make it difficult to search across various domains. For example, would a scholar need literature from economics and agriculture, he or she would have to access two different DLs, relying on the heuristic manner, which often requires step-wise or extensive navigations through the affected DLs. Achieving interoperability by crosslinking publications

from different repositories is still an open field of research when the interconnection between different domains is tackled [Dors17, Jacs05].

The current practices of Google Scholar, BASE - Bielefeld Academic Search Engine, Mendeley or Semantic Scholar from AI2, and many more discovery systems that achieve to integrate and index hundred millions of metadata sets from different libraries, are impairing the barriers by fading the isolation aspect of repositories. However, challenges such as crosslinking resources¹, i.e., scientific publications with an assured degree of semantic similarity remain even today. That issue certainly presents a complex process of lexical or string matching, mostly due to the diversity of ontologies and metadata vocabularies used for describing resources [JJHY12]. Hence, retrieving and recommending publications from these repositories continues to rely on the metadata terms rather than on vocabulary, i.e., thesauri alignments between repositories.

The usage of linked open data, i.e., the aligned concepts between repositories, can be seen as hope for breaking down the heterogeneity among repositories. Linked Data (LD), as a way to publish data in a structured format, has achieved to enhance the meaning and usability of data by establishing interlinks between them, across repositories. Therefore, in addition to the links between the documents, the links between the data, in even the finest granularities, make it possible not only for people but also for machines to query and create knowledge of the data. This empowers the usage of ontologies as models for formal representation of taxonomies, classifications, and relations among the data, concepts, or entire entities. Hence, nowadays there are several such vocabularies designed for various data descriptions, such as Friend Of A Friend (FOAF) for people in social networks, Dublin Core (DC) for digital resources, Simple Knowledge Organization System (SKOS) for representing KOS on the LD, etc. Given the recommendations for the use of existing vocabularies, in contrast to the creation of new vocabularies within the repositories, it affects the reduction of heterogeneity and the increase of interconnection between them. Chapter 3 provides more details about linked data and semantic technologies. In addition, the introduction of LOD emphasized even more the role of thesauri, especially in the mapping process, i.e.,

¹ Appendix A1.1 provides a definition of the “resource”.

1. Introduction

aligning together concepts from different repositories that bear the same meaning. In this way, the concept "Inflation" from the STW thesaurus [Stw17] is mapped to several other thesauri/vocabularies (WIKIDATA, AGROVOC, TheSoz, DBpedia, GND, JEL). As a result, each publication described by the STW concept is connected to all the other publications where descriptions from mapped KOSs are used. By considering the SKOS navigating hierarchy with the related, broadened, or narrowed concepts, the interconnection of concepts becomes more inclusive. Section 6.1 provides more details in regard to our approach to the application of thesauri and term alignments.

In the process of generating recommendations, namely finding related publications for a selected publication, the combination of terms from metadata has a decisive role. In discovery systems, we are indeed dealing with an enormous amount of data, but what emerges is the lack of functionalities for making refinements and adjustments of particular metadata components in order to narrow down the results. For example, by considering Google Scholar, there is an evident limitation in the number of facets for further filtering and thesauri - or disciplinary based - searches. Moreover, if we are interested in emphasizing or diminishing the role of a particular concept during the search, then it becomes even more difficult. Let us assume that we have found an interesting publication in our favorite DL, entitled "Globalization, brain drain and development". If we prefer to get a list of recommended publications related to "brain drain" rather than "globalization", such adjustments are very necessary.

Referring to current search practices, when a new search is initiated or expanded in a DL, it is principally based on keywords, i.e., the user input of terms for articulating the query. Thus, the effort of the user to choose the right terms is excessive and may be iterative. Hence, we are proposing an automatic approach for extending or enriching the provided terms with concepts from controlled vocabularies, including terms generated by machine learning techniques, in a way to facilitate the searching process and reduce the mental workload. Moreover, such enrichments may lead to further discoveries, thus finding publications that may be in your interest in an unintended manner. Such approaches, by proposing specific cases, are set out in section 6.4.

Another very important element, which is evident almost in every DL, deals with the exact identification of publications that belong to a particular author. Hence, relying only on the author's name it is very difficult to harvest all the research output of a given author from a particular repository. Even more, this can be considered as impossible if we attempt different repositories. What makes the process challenging is the appearance of an author with different name alternatives, inside one or across repositories. Moreover, by considering the probability of having different authors with the same name, the complexity becomes obvious. Therefore, crosslinking information based on a particular author or co-authorship relation has to require the author disambiguation and the application of already known persistent identifiers. Currently, there are many efforts in that direction, by crosslinking and extending author identifiers. Such an example is the FREYA project for interconnecting identifiers in a way to improve the interoperability of data [WiFe18]. The Scholia Web service, for generating on-the-fly scholarly profiles, with several other functionalities such as co-author, topic and citation graphs, is another promising use case [NiMW17]. The Scholia Service collects the data by querying the WIKIDATA SPARQL endpoint, hence the community input can be of significant importance.

The application of such identifiers already is part of several DLs. Based on what we have observed, the level and quality of their deployments, i.e., author disambiguation, is different. Thus, in several cases, DLs are facing an entirely ambiguous author set, which means none of the authors are identified with any type of a local or global identifier. Therefore, the only possibility for retrieving author's research output relies on her/his name, and that typically results in very low precision. In other DLs, merely a part of the collection contains identifiers for authors. Such a partial disambiguation state currently is evident at EconBiz². Thus, from a total of around 10 million bibliographical records, i.e., publications, there are around 500 000 authors that are identified with a persistent identifier (GND ID in this case). Concerning the rest of the authors, an additional clustering and disambiguation process must be followed. Accordingly, we have proposed an approach, as presented in section 7.2, that can be applied in

² <https://www.econbiz.de>

1. Introduction

either entirely or partially ambiguous environments, assigning a globally known persistent identifier to authors, such as VIAF ID.

Thus, by having a persistent identifier, we can not only generate a reliable list of publications that surely belong to that author, but an extended profile can also be created by attaching different library and non-library resources, based on the LD approach. Having in mind the potential of WIKIDATA, it can serve also as a hub to extend the list of identifiers. The outcomes of such enrichment and profile generation are described in section 9.2.

The achievement of such interoperability among DLs by crosslinking publications, authors and other related data would facilitate scholarly communication, scientific findings, knowledge retrieval, and representation. Starting from a single point of access, a scholar would be able to find publications and authors, previously enriched with additional information, from different repositories.

1.2 Overview of Approaches and Contributions

The work presented here focuses on the process of crosslinking scientific publications stored in a specific repository with related data, such as publications, author information, correlations with other authors, information about conferences, events, etc. For this purpose, the thesis pursues two ways of retrieving the most relevant information from the targeted repositories³, which will be explained below. Both crosslinking approaches start from one repository of a DL. While in the first case the crosslinking process begins based on publication-centered metadata, the second takes the author metadata as its point of departure. The process then proceeds by interlinking information among several repositories.

Following the first approach, the content of repositories published based on the semantic web technology stack, such as bibliographic linked open data and authority linked open data repositories [BHIB08], are among the

³ Appendix A.1 provides details about the meaning of “target repository”.

1.2 Overview of Approaches and Contributions

first where the deployment of these strategies will be evaluated. By bibliographic repositories, we mean repositories that contain metadata about publications, books, authors, or any digital content. As authority repositories, on the other hand, we consider repositories that contain authority name information concerning persons, such as, name alternatives, cross references, useful for identifying and clustering an author. According to the first approach, the interlinking of scientific publications primarily relies on existing alignments of concepts between KOSs used to index resources in repositories. Regarding this idea, we define the first research question.

RQ1.1. How could existing methods from information retrieval, relying on terms alignments, be extended or combined to retrieve similar publications from different repositories?

The exploration of RQ1.1 investigates whether utilizing term alignments between repositories is helpful for retrieving semantically similar publications. Moreover, it emphasizes the role of thesauri in the source and target repository and the implications of not using thesauri. In order to answer this RQ, the thesis explores different scenarios on the basis of the experimental results and seeks to determine which one works best (for example, the use of alignments between repositories or thesauri for retrieving a preliminary subset of possibly similar publications).

Extending with similar publications from across repositories brings new challenges to the users. In this case, they face too many choices to select from – a situation of information overload – which needs to be addressed. In handling information overload-related challenges, we rely on the semantic similarity measure, data mining, and machine learning techniques. In Chapter 6, respectively in section 6.1, a detailed description of this question is given.

RQ1.2. How can machine learning methods improve the quality of retrieved publications from different repositories?

RQ1.2 implies that the application of the data approaches will be used to measure the semantic similarity or relatedness between publications retrieved from different repositories, especially from repositories of a different discipline

1. Introduction

than the initial⁴ one. Namely, being in a DL of the economic domain, we will be able to retrieve publications from social, medical, or agronomic domains. Even so, the results generated by such measurements can serve us for the rankings of retrieved publications. For this purpose, we intend to follow two different approaches. Initially, we measure the similarity among publications in a very traditional way, by applying the algorithms for the Vector Space Model (VSM), and follow with the most comprehensive approach that is proclaimed today, namely the word embedding (WE) approach. Thus, the vector representation of words using neural networks, i.e., word embedding, is applied in the same context. Comparatively, the evaluation of both approaches will assess which one suits best in particular circumstances. The approaches are elaborated in sections 6.2 and 6.3, while the assessments are part of chapter 8.

In order for the approaches above to apply and operate between publications, apart from the concepts between the alignments, it will be necessary to include more information, i.e., the metadata of publications. For this reason, the inclusion and the selection of different elements from the data that describe a publication is a significant factor. As a result, the question that naturally follows is:

RQ1.3. What methods from text mining and natural language processing should be extended, combined or adapted, to determine the key terms of a publication which are suited for a semantic similarity between publications?

The outcome of this question is of particular importance as it is a prerequisite for applying other approaches, especially data mining-related approaches. More details of this point are given in section 6.2.1.

Another issue that can be addressed at this point is the use of linguistic thesauri, such as WordNet, which contains an extension of terms and their respective synonyms. Therefore, we are interested to figure out the implication of external resources, such as WordNet synonyms, for particular terms in the existing set of data, for improving the similarity degree among initial and retrieved publications.

Given all these approaches and components, we intend to propose an applicative solution that could include all of this in a single interface.

⁴ Appendix A1.1 provides details about the meaning of “initial repository”.

1.2 Overview of Approaches and Contributions

Starting from the goal to generate and enrich author profiles with data not found in the initial repository⁵, our approach propagates crosslinking author information with library and non-library resources from several repositories. Thus, the scholar would be able to find in one place additional publications, co-authors, biographical, and other information related to that author. To achieve this goal, the presence of persistent author identification attributes may be decisive for the accuracy and quality of the retrieved information. Therefore, the research question in this context is:

RQ2.1. What methods to harvest author-related information exist, particularly for cases where the author does not have a unique global identifier?

Name ambiguity is a real and persistent problem in the world of DLs. In many situations, we face the disambiguation challenge for authors with the same name, or when the name of an author is presented in different variations. Hence, in many cases, it is difficult or almost impossible to decide whether a particular research output belongs to a specific author or not. The presence of author identifiers somewhat alleviates this challenge. We address RQ2.1 in chapter 7, while the main outcomes of such interoperability are highlighted in chapter 9. In this RQ we consider two cases with regards to author identification:

- In the absence of an identifier, harvesting author information from other repositories becomes more difficult, since we face author name ambiguity scenarios. Therefore, the process should go through the author name disambiguation workflow.
- When the author in a DL or repository is identified with a global identifier such as GND, VIAF, RePEc, ORCID, and so on, finding and retrieving data from other repositories can be made with greater fidelity. However, the diversity of these factors also causes difficulties in the identification process. For example, in repository A the author can be identified with a GND identifier, while in repository B with a RePEc identifier. This requires us to extend the range of the identifiers to match the search with the identifier in the repository from where harvesting will take place.

⁵ Definitions of terms, such as initial and target repositories are given in Appendix A.1.

1. Introduction

Apart from examining different approaches for crosslinking publications and authors, the thesis contributes by analyzing and implementing several methods in the domain of information retrieval and recommender systems. The application of these methods, by representing documents as VSM and the document vocabulary representations through Word Embedding (WE) methods, is done comparatively. In this way, through different scenarios, we emphasize the advantages and limitations of such methods, especially in regard to their applicability in the domain of DLs. We propose an integration and combination of several methods to improve the quality of the retrieved publications, i.e., getting the list of semantically related publications. In addition, the usage of external thesauri and concept alignments in the context of terms enrichment and further performance enhancement is applied. Concerning the user experience and evaluations, we have deployed an interface that integrates the proposed methods with the possibility of their adjustments and customization.

A distinct contribution is also given to the process of author disambiguation by proposing and applying an algorithmic approach to this purpose. The approach can operate in different environments, i.e., in partially or entirely ambiguous repositories, relying on services outside the repository, such as VIAF and WIKIDATA. Consequently, linking and collecting data on authors becomes possible and leads to the creation of a comprehensive profile.

1.3 Publications

Parts of this thesis were already published in the following research papers:

- Arben Hajra, Tamara Pianos, Klaus Tochtermann. 2021. “Linking Author Information: EconBiz Author Profiles”. In: *Metadata and Semantics Research, MTSR’20*. Madrid, Spain. pp. 180-191. CCIS, vol. 1355, Springer, Cham. DOI: 10.1007/978-3-030-71903-6_18.
- Arben Hajra, Klaus Tochtermann. 2018. “Visual Search in Digital Libraries and the usage of External Terms”. In: *22nd International Conference Information Visualisation (IV)*. Salerno, Italy. pp. 396 - 400. IEEE. DOI: 10.1109/IV.2018.00074.

- Arben Hajra, Klaus Tochtermann. 2017. "Linking Science: Approaches for linking scientific publications across different LOD Repositories". In: *International Journal of Metadata, Semantics and Ontologies*. vol. 12, No. 2/3, pp.124-141. DOI: 10.1504/IJMSO.2017.090778.
- Arben Hajra, Klaus Tochtermann. 2016. "Enriching Scientific Publications from LOD Repositories through Word Embeddings Approach". In: *Metadata and Semantics Research, MTSR'16*, Göttingen, Germany. pp. 278-290. CCIS, vol. 672, Springer, Cham. DOI: 10.1007/978-3-319-49157-8_24.
- Vladimir Radevski, Arben Hajra, Fidan Limani. 2016. "Semantically Related Data as Technology-Enhanced Support for Research Assistive and Quality Tools". In: *Technology Advanced Quality Learning for ALL, QED'16*. Sofia, Bulgarian National Commission for UNESCO. pp.18-31. ISBN: 978-619-185-260-4.
- Arben Hajra, Vladimir Radevski, Klaus Tochtermann. 2015. "Author Profile Enrichment for Cross-Linking Digital Libraries". In: *19th International Conference on Theory and Practice of Digital Libraries, TPD L 2015*. Poznań, Poland. pp. 124-136. Lecture Notes in Computer Science, vol. 9316, Springer, Cham. DOI: 10.1007/978-3-319-24592-8_10.
- Arben Hajra, Atif Latif, Klaus Tochtermann. 2014. "Retrieving and ranking scientific publications from linked open data repositories". In: *14th International Conference on Knowledge Technologies and Data-driven Business, (i-KNOW '14)*. Graz, Austria. pp. 1-4. ACM, New York, USA. DOI: 10.1145/2637748.2638436.
- Arben Hajra, Klaus Tochtermann, Vladimir Radevski. 2013. Enriching scientific publications with semantically related data. In: *BCI'13 Proceedings*, p. 140. BCI, CEUR-WS.org, vol. 1036, Thessaloniki, Greece. urn:nbn:de:0074-1036-1.
- Arben Hajra, Vladimir Radevski, and Atif Latif. 2012. "Enhancing Scholarly Communication Results and Search Experience by Interlacing Relevant Scientific Repositories". In: *Proceedings of the 7th Annual South-East European Doctoral Student Conference*, pp. 517-527. SEERC, Thessaloniki, Greece. ISBN: 978-960-9416-05-4.

1. Introduction

1.4 Chapters Structure

This thesis is structured in four main parts and an appendix.

Part I covers the foundations of the thesis, general information of the domain, terms and concepts, technologies, services, algorithms, and repositories used in this work.

Chapter 2 presents an overview of Scholarly Communication with a particular focus on Digital Libraries (DL) and interoperability among resources.

Chapter 3 highlights the Semantic Web Technologies and Linked Open Data (LOD), especially bibliographic repositories and thesauri offered as LOD. The alignments between these repositories are also elaborated. The presence of Integrated Authority files, such as VIAF or WIKIDATA, is considered as a hub for integrating and crosslinking authors.

Chapter 4 gives details about Recommender Systems and techniques for measuring the semantic similarity degree among text corpora, starting with the classical Vector Space Model (VSM) and continuing with Word Embedding (WE) approach.

Part II contains the main contribution for crosslinking or enriching scientific publications.

Chapter 5 is dedicated to the main idea and followed approach for crosslinking and enriching a DL resource with other information. Details about the publication- or author-centered metadata are presented, on both sides, at the initial and targeted repositories.

Chapter 6 explains the approach for crosslinking information regarding LOD Repositories. It begins by exploring the existing alignments among repositories, and continues by evaluating the text-mining techniques for achieving improvements about the semantic measurements between resources. Two main approaches are followed for this purpose, the Vector Space Model through TF-IDF and Cosine Similarity, in comparison with the Word Embedding approach through Word2Vec. The chapter ends by introducing user interactivity for deploying both of them in a single search interface.

Chapter 7 focuses on the approach for crosslinking information concerning author metadata. The presence of persistent author identifiers is of crucial importance, therefore WIKIDATA is considered as a hub for further expansions. In the absence of any identifiers, the process goes through author name disambiguation, where the usage of VIAF is considered. We introduce an algorithmic and formal approach for the author's identification at VIAF.

Part III is about the evaluation of implemented approaches.

Chapter 8 gives and discusses results regarding the usage of LOD alignments and data mining methods for crosslinking resources. The application of the Vector Space model and Word Embedding approach is evaluated comparatively.

Chapter 9 represents the results as the output of author-related information. In this chapter, discussions and outcomes are focused on enriching author data through identification, namely the author name disambiguation.

Chapter 10 represents the related work. Related work is positioned at the end of the thesis for the following reasons. Initially, most of the approaches, methods, and algorithms mentioned in related work require a previous explanation of the problem as well as the context for what we are referring to. Hence, having it at the beginning may cause interruptions to the flow. Moreover, placing related work just before the conclusion interlinks the existing approaches with our summary and key findings. However, a considerable part of the related work is also cited in the relevant section in the corresponding chapters.

Part IV represents the conclusion and future work.

Chapter 11 concludes the thesis.

Chapter 12 presents the future work.

Part I

Foundations

Scholarly Communication

“panta rhei...”

HERACLITUS

Libraries present the first and foremost source for scholarly communication. Traditionally, they provide the basic information infrastructures, the content and the metadata for sharing and discovering knowledge. They can be categorized as a primary source from where scholars are provided with resources.

2.1 Digital Libraries

During the era of digitalization, libraries have become an even more crucial primary source of scholarship, by increasing and simplifying the accessibility of resources [BoFu02, KIMc99]. According to Rowley and Hartley[RoHa17], a digital library can be viewed as a managed collection of digital information with associated services, accessible via network. Thus, at present, they can be categorized in different levels, such as national, institutional/university, or domain-specific libraries. For example, the German National Library (DNB), the Library of Congress (LoC) and the British Library (BL) represent examples of national DL for Germany, the US, and the UK, correspondingly. Furthermore, institutional repositories seize and preserve the research output of single or multiple institutions by providing a component that increases access to research and competition, brings economic relief, as well as reduces the monopoly character of power journals [Crow02]. At the same time, they can serve as indicators of a university’s quality, by increasing the visibility, status, and public value of the institution [Crow02]. Among the large number and

2. Scholarly Communication

variety of such repositories, the Smithsonian/NASA Astrophysics Data System (ADS), Kyoto University Research Information Repository, MIT Institutional Repository, or the CERN Document Server, are the most popular at the moment. The open-source system DSpace is one of the most favorite platforms for developing and maintaining institutional repositories [TBSB03]. Finally, the Sterling Memorial Library at Yale University or the Baker Library at Harvard Business School are but few examples of institutional/university DLs.

The thematic division is quite common in the world of digital libraries. Therefore, nowadays there are several domain-specific DLs, such as economics, medical, social sciences, computer sciences, etc. Such examples include the National Library of Medicine (NLM) in the United States, the German National Library of Medicine (ZBMed), Leibniz Information Centre for Science and Technology (TIB), etc.

The role of DL is undisputed in the overall scholarly communication process. With their global and simplified accessibility of resources, their usages bring a huge benefit for the community and scholars in particular. However, not always does the DL satisfy each scholar's request. In some cases getting the most relevant and qualitative resources in a reasonable time using a DL can be challenging [ASWF14, BoFu02, Borg10, Than14].

Publications stored in a repository in most cases belong to a particular domain, described or cataloged according to predefined metadata schema, by trained professionals in the field of library/information sciences. This practice leads to some limitations in the search of literature from a specific field, based on the applied cataloging and indexing rules. Such that many studies are categorizing DLs as monolithic systems, where metadata describes the data rather than uses [Borg99, Tenn04]. Therefore, the MARC (MACHine-Readable Cataloging) format, with all its variations, does not offer almost anything considering the relationship between data, especially the data outside any repository [AISR12, Tenn02].

By triggering a publication in a particular DL, apart from the standard metadata used for describing that publication, the system can offer some metrics such as downloads, views, citations and a list of related publications stored in that repository. However, do scholars need more? What about related publications stored or indexed in different libraries, new author correlations and other important information for enriching that resource?

2.2 Interoperability of Digital Library resources

The interoperability between DLs has been a central concern from the beginning of their creation. As it is mentioned in [PCWG98], researchers have been struggling with interoperability as one of the main features for achieving better recommendation results from digital libraries.

2.2 Interoperability of Digital Library resources

The role of the current DLs is more than evident; however, there are several directions where they lack to provide the needed service. One of the most obvious shortcomings is the need for a proper link between resources in different repositories, i.e., the visibility and accessibility of a resource stored in a repository from different DLs. As mentioned in the motivation part, repositories are considered as isolated silos. Therefore, it is almost impossible to harvest resources from different repositories with the same query formulation. The difficulty is mainly due to the diversity of ontologies and metadata vocabularies used for describing resources [JJHY12], including the domain-specific information. Searching through cross-disciplinary repositories (e.g. economics, agriculture, medicine) makes it necessary to perform a particular search in several places. All this is still very heuristic, and often requires step-wise or, as far as possible, simultaneous navigations through the affected DLs.

The interoperability among resources has been represented as a problem for many years [Bess02, Borg02, PCWG98, Shet99] and continues to be the subject of research until today [AgFS16]. Consequently, according to Agosti et al. [AgFS18], nowadays DLs started to be perceived as user-centered systems, versus the document-centric approach that was a characteristic of traditional libraries. Therefore, the vision of DLs changes in many aspects, where, among the others, the management of resources now is considered as a collaborative task. Thus, they managed to improve the utilization of resources from different perspectives, such as quality of services, system performance and user experience [GaGF10, HFCH12, Xie06]. Therefore, the isolating character of DLs must be something that needs to be overcome or at least minimized to the extent possible.

2. Scholarly Communication

The achievement of interoperability among DLs by crosslinking publications, authors and other related data would facilitate scholarly communication, scientific findings, knowledge retrieval and representation [Than14]. Starting from a single point of access, a scholar would be able to find resources, i.e., publications and authors, previously enriched with additional information from different (disconnected) repositories.

2.3 Integrated Authority Files

Researchers, i.e., author records, are part of several DLs and other services that index their published works. However, not all of them have adopted a unique way to represent an author's name. Therefore, the same author may be in different name variations inside or across several repositories and services, known as name synonyms. For example, William Nordhaus is represented with several spelling alternatives, such as W. D. Nordhaus, U. Nordchautz, W. Nordhaus, and Weilian Nuodehaosi. Besides, different authors may generate research outputs under the same names, i.e., name homonyms. This represents one of the main obstacles for linking authors' profiles between different repositories.

Nowadays, there are several efforts for generating authority profiles for aggregating and uniquely identifying resources and authors. Consequently, for each author, a particular profile is generated and a global persistent identifier is assigned. In this way, the interlinking process among these services would be simplified, especially if a particular repository (service) offers outgoing links to other repositories. Hence, the data exchange among repositories would be possible, and the profile of an author would be always up to date in all of them, smoothing out the effect of isolated repositories.

As most applicable approaches that are operating in the area of author disambiguation or used as a "hub" for retrieving accurate information from other repositories we emphasize: ORCID, VIAF, ISNI, VIVO, Google Scholar, Scopus, Mendeley, ResearchGate, Academia.edu, arXiv, Microsoft Academia.edu, ResearcherID, and OpenID. Some of these services, such as Academia.edu or ResearchGate, are more oriented to social networking among researcher communities. In such services, including several others,

2.3 Integrated Authority Files

i.e., ORCID, ISNI, RePEc, author contributions can be of huge impact for identifying themselves and their research outputs. Several other services are focused on completely automated approaches for clustering and disambiguating authors. Some of these approaches we have described in our previous paper [HaRT15], and with several supplementary details are presented as follows:

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 LoC, DNB BNF and OCLC [HiTo14, Loes11].

GND - The Integrated Authority File (GND-Gemeinsame Normdatei) is an authority file for persons, corporate bodies, conferences and events, geographic information, topics, and works. Above all, it is used for the cataloging of literature by libraries, but it also is increasingly deployed in archives, museums, projects, and web applications. It is operated cooperatively by the German National Library, all German-speaking library networks, the German Union Catalogue of Serials (ZDB) and numerous other institutions. Contributions to the GND are made either via the networks or in direct agreement with the German National Library. GND is one of the biggest contributors to the Virtual International Authority File (VIAF), among the other national authority files [Dnb16]. By querying the offered dump files, currently, GND results to have 4 917 517 differentiated persons. From that number, 4 896 088 (99.6%) contain VIAF ID in their authority records, while 5 836 have ORCID ID.

ISNI - International Standard Name Identifier is a registry providing reliable identifiers for public identities including persons and organizations. Just these identifiers are considered as a key element for facilitating and making possible the data interlinking process among repositories, i.e. digital libraries. Similarly as VIAF, ISNI currently is maintained by OCLC. Even though their goals converge at a point, there are also changes in the way of organization and functioning. Therefore, the VIAF is using selectively the ISNI data for the cluster's correction and enrichment. Hence, an ISNI identifier can be noted in the VIAF clusters also [MaAG13].

2. Scholarly Communication

ORCID - Open Researcher and Contributor Identifier create and maintain a registry of unique researcher identifiers and a method of linking research activities. The main contributors are several publishing houses, scientific communities and universities. It has available APIs under an open-source license [Haak13]. At the moment, ORCID can be characterized as a very popular service for identifying authors. A large number of institutions, conferences or magazines recommend or even make compulsory the use of ORCID IDs for authors. Furthermore, other authority files, i.e. GND, started to enrich their bibliographical records with ORCID detail [HaPa17].

VIVO - enables the discovery of researchers across institutions. It is an open-source semantic web application, where institutions such as Cornell, Harvard, and Indiana University, manage and publish information about researchers and their activities⁶.

ResearcherID - to identify potential collaborators and avoid author misidentification, each member is assigned a unique identifier to enable researchers to manage their publications' list. The ResearcherID integrates the data with the Web of Science of Thomson Reuters Company and ORCID⁷. From April 2019, ResearcherID identifiers claimed publication history and other ResearcherID account information will be moved to Publons⁸.

OpenID - is a foundation that promotes Open ID technologies. OpenID Foundation members include leading companies and individuals in the digital identity industry such as Google, Microsoft, and Yahoo. Even though this currently has no direct application in scholarly communication, however, there is a promising potential for Internet-scale user-centric identity infrastructure [ReRe06].

RePEc ID - The RePEc short-ID is a permanent identifier that is uniquely assigned to people, mainly from the field of economics. RePEc is a noticeable example of showing the efficiency of a service when the input of authors and publishers is evident [KrZi13].

⁶ What is VIVO?, <https://duraspace.org/vivo/about/>, accessed 07.09.2018

⁷ What is ResearcherID, <https://www.researcherid.com/>, accessed 08.09.2018

⁸ <https://publons.freshdesk.com/support/solutions/articles/12000055561-what-is-happening-to-researcherid->, accessed 27.06.2019

2.3 Integrated Authority Files

WIKIDATA – is a free, multilingual open knowledge base that can be read and edited by both humans and machines. It acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others [Wiki18]. Currently, the knowledge base contains more than 69 million items, where more than 6 million are related to humans. Section 2.3.2 provides more in-depth information about WIKIDATA.

In this work, we consider VIAF and WIKIDATA as resources with the most usage relevance. The approach of clustering authors through the discovery and dissemination process by integrating authority files from several other DLs is the main reason for the VIAF selection. While WIKIDATA, being one of the most important hubs for crosslinking several identifiers, i.e., authority data, where community contribution is essential, represents an excellent opportunity to find and use author-related data.

2.3.1 Virtual Authority Files (VIAF)

The current practice followed by several national libraries to create and maintain authority files on their own brings a distinctive way of preserving them [HiTo14]. VIAF, on the other hand, aims to link and combine authority files from several national libraries into a single “super” virtual authority record, i.e., cluster. Therefore, VIAF is offering a freely available API that can be used by anyone without the need for authentication. In addition, the VIAF LOD repos are another alternative to the API access. However, VIAF strongly recommends using the API for up-to-date information, according to the frequency of updates.

VIAF links different name formulations for the same person by integrating authority files from more than 40 contributors (national libraries and institutions), from more than 30 countries⁹. This number increases continuously, as new contributions become part of VIAF clusters. One of the biggest contributors to VIAF is the DNB - German National Library. Besides national libraries, VIAF is also focused on other sources such as Getty ULAN, WIKIDATA, Perseus, Syriac, and xR.

⁹ <http://www.oclc.org/viaf.en.html>, accessed 23.08.2018

2. Scholarly Communication

The aggregated sources are clustered and identified with a globally unique identifier, i.e., a VIAF ID. However, there are cases when the VIAF clustering algorithm exposes various issues, such as numerous clusters for the same person, different sources (different people) into the same cluster, incorrect bibliographic data, or clusters with poor content (lack of information). The aggregation of sources inside a particular cluster has an accuracy of around 99% (see the publication “Managing Ambiguity In VIAF” [HiTo14]). Therefore, according to this resource, if two sources have less than a 1% chance of describing the same person, they are excluded from that cluster. Therefore, it may be possible for the same person to have more than one cluster. Based on the outcomes from [FWMJ12], a search of 283 114 different name labels resulted in 59% unambiguous output, meaning that only one heading cluster was retrieved, 26% matched two clusters, 10% matched three clusters, 3% matched four, and 2% more than four. As a result of changes that may occur inside of a cluster, such as new titles, co-authors, or author details, authority records may be moved from one cluster to another. Therefore, searching at different times can result in different results. The re-clustering frequency is monthly.

The VIAF data consumption can be done in several forms, such as the simple search and advanced (SRU-based) search at viaf.org, through the API usage, or by downloading the dump files at viaf.org/viaf/data. The API makes it possible the search for authority data by keywords, names, title, etc., while the dump files provide data about clusters, external links to other resources and even internal links between clusters, in case of merge/split.

Based on the 2016 statistics [Hick16], there were 55 million source authority records, 130 million bibliographic records, 256 million links between sources, 30 million external links, and 33 million VIAF clusters.

2.3.2 WIKIDATA

Each item in WIKIDATA, including persons, is uniquely identified with a number, preceded with a “Q”. For example, the WIKIDATA identifier of the American economist “James Heckman” is Q312561. The structured data in the WIKIDATA repository are described through the property value pairs called statements [VrKr14]. The properties always have the prefix “P”

2.3 Integrated Authority Files

followed by a specific number. For example, P227 is the property for the GND ID.

As mentioned in [VrKr14], one of the most essential factors in WIKIDATA development is the volunteer community's reuse and integration of external identifiers from existing databases and authority controls. These external persistent identifiers allow applications to integrate WIKIDATA with data from other sources that remain under the control of the original publisher. Accordingly, WIKIDATA and VIAF represent one of the most important hubs for interconnecting authors' identifiers [Neub17].

In the following, we have analyzed the presence of the most significant authority identifiers in WIKIDATA. Table 2.1 provides more details from a statistical point of view comparing a one-year period, 2019 versus 2018, and by including here the data from September 2020. As shown, there is a noticeable increase in almost all identifiers. From around five million people on WIKIDATA in 2019, 1.2 million are identified with VIAF, while 613 000 with GND ID. Furthermore, if these figures are compared with the data in 2020, we see more than their doubling. The increasing presence of GNDs and ORCID identifiers in WIKIDATA within a year also emphasizes the importance of these identifiers in the community. Moreover, from the WIKIDATA perspective, it is worth mentioning that the GND identifiers are almost completely attached to VIAF as denoted in the table (in February 2019, from 613 051 GND identifiers 611 478 are mapped to VIAF, see VIAF+GND).

Table 2.1 The list of some authors' identifiers in WIKIDATA

	Feb 2018	Feb 2019	Diff %	Sept 2020
WIKIDATA(human, Q5)	4 128 338	4 887 509	18.39%	8 162 753
VIAF (P214)	1 013 751	1 188 858	17.27%	2 558 168
GND (P227)	496 960	613 051	23.36%	1 023 323
ORCID (P496)	100 760	432 384	329.12%	1 602 716
RePEc (P2428)	6 829	6 849	0.29%	6 942
VIAF+GND	495 344	611 478	23.45%	1 011 227
GND+RePEc	4 344	5 635	29.72%	6 165
GND+ORCID	1 187	2 164	82.31%	1 0037

Semantic Web and Linked Open Data

*"Invisible threads are
the strongest ties."*

FRIEDRICH NIETZSCHE

The Semantic Web aims to improve the current state of the World Wide Web [AGHH12, BHLO01] - not by offering an alternative to the current Web but by offering an attempt to extend it. The main aim is achieving that the data on the Web to become machine-understandable information, independently of platforms and other boundaries. The Semantic Web provides the technologies and standards that are needed to add machine-understandable meanings to the current Web, thus computers can understand the Web documents and therefore can automatically accomplish tasks [HiKR09, Yu11].

The implementation of semantic technologies and the interoperability between different linked data sources are considered for enriching digital libraries with additional information.

3.1 Linked Open Data (LOD)

Linked Data has been introduced and conceptualized by Tim Berners-Lee as a set of best practices for publishing and interlinking structured data on the Web [BHIB08, BiHB09, HeBi11]. Linked Data is about employing the Resource Description Framework (RDF) and the Hypertext Transfer Protocol (HTTP) to publish structured data on the Web and to connect data between different data sources, effectively allowing data in one data source

3. Semantic Web and Linked Open Data

to be linked to data in another data source [BeMc04, BHIB08, BiHB09, HeBi11].

The Resource Description Framework (RDF) was proposed as a model, similar to Entity-Relationship (ER), for processing metadata and providing interoperability between applications that exchange machine-understandable information on the Web [BeMc04, CyWL14, LaSw99]. A basic element of RDF and Semantic Web construction is the statement. It represents the triplet, resource together with its property and the value for that property. These three elements of a statement are known as subject, predicate, and object. The proposed syntax for serializing RDF is XML in the RDF/XML form, designed for machine consumption rather than for human eyes. There are indeed other RDF serialization formats, such as Notation-3 (or N3), Turtle, and N-Triples [CyWL14]. In RDF, each statement or triple represents a single fact. A collection of statements or triples, which form a graph, represents some given piece of information or knowledge.

The other levels above RDF consist of vocabularies for describing properties and classes of RDF resources, such as RDF Schema (RDFS) and Web Ontology Language (OWL). Resource Definition Framework schema (RDFS) allows users to define their own terminology, i.e., vocabulary for representing RDF statements. RDFS describes the relationships between objects by creating hierarchies of classes and properties [BrGM14]. "Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms" [Fens01, MaSt01]. For more complex ontologies, where it is necessary the use of several vocabularies, the deployment of OWL offers an extended construct over RDFS [McHO04].

Below we elaborate three basic n-triples regarding a particular author in a given repository. There should be noted that different repositories may use different ontologies and vocabularies for representing the same thing. (e.g., *dc:creator*, *foaf:maker*). There are evident cases where data publishers apply their own ontology, again the strong recommendation for using the already existing ontologies and widely deployed vocabularies.

```
<http://linkeddata.econstor.eu/beta/resource/authors/9060227> foaf:name  
"Kehl, Victoria".
```

```
//The author with number 9060227 is called "Kehl, Victoria".
```

3.1 Linked Open Data (LOD)

```
<http://linkeddata.econstor.eu/beta/resource/publications/30811> rdfs:label
"Identification of responders to Amiodarone subgroup analysis of the EMIAT study".
//The publication with number 30811 is titled "Identification of responders to Amiodarone
subgroup analysis of the EMIAT study".
<http://linkeddata.econstor.eu/beta/resource/publications/30811> dc:creator
<http://linkeddata.econstor.eu/beta/resource/authors/9060227>
// The author of this publication (30811) is "Kehl, Victoria" (9060227).
```

The development of the Semantic Web appears as a layered process with layers that interact between them. Figure 3.1 represents a comprehensive representation of the Semantic Web architecture, unlike several variations of Tim Berners-Lee Semantic Web LayerCake [Bern00].

After the designing process (i.e., defining RDF, RDFs or OWL) there is a possibility to retrieve information by querying RDF data. In this context, the available tool is the Simple Protocol and RDF Query Language (SPARQL) [HaSe13]. As RDFS and OWL describe properties and classes in RDF, in a similar way SPARQL can be used for querying ontologies, i.e., knowledge bases, and diverse data sources directly. SPARQL is also a protocol for accessing RDF data and not only a query language [PrSe08].

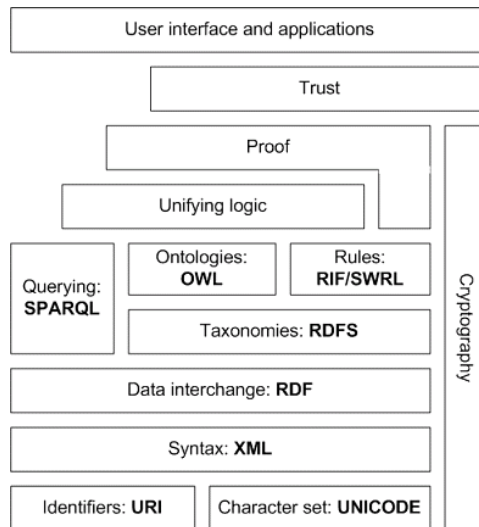


Figure 3.1 Semantic Web architecture in layers [Brat07]

3. Semantic Web and Linked Open Data

As mentioned in [FMFG18], the use of RDF to expose semantic data has seen a dramatic increase over the last years, making RDF data ubiquitous. Therefore, as of May 2007, there have been evident only 12 datasets, comparing to 1 184 datasets counted in April 2018. Figure 3.2 gives an overall view of the current Linked Open Data Cloud Diagram and interlinks with other datasets in the cloud (the current version has 15 993 links). The size of the circles matches the number of edges connected to each dataset. Thus, in total there are three sizes, large with more than 100 edges, medium 50-100 or small with 1-5.

The color coding in the figure denotes the different domains, such as Life Sciences, Government, Geography, Linguistics, Media, Publications, Social Networks, User Generated and Crossdomain datasets. It is worth mentioning that various datasets are published under a specific license or as public.

Among others, our interest is focused on the datasets from the publications' domain. The diagram shows that they take an important part in the cloud, with around 10% in total based on 2014 statistics. Thus, inside these datasets, the metadata or the entire catalog of the offered data can be found. Some of the libraries one the LOD of interest for our scenarios are German National Library (DNB), Library of Congress (LoC), British National Bibliography (BNB), Swedish National Library (LIBRIS), Hungarian National Library (NSZL), Europeana Digital Library, Leibniz Information Centre for Economics (ZBW), Computer Science Bibliography (DBLP), Multilingual Bibliographic Database for Agricultural (AGRIS), Association for Computing Machinery (ACM), etc.

3.2 Selected Repositories

In the following, we list the repositories used for developing and evaluating our approaches. Initially, the experiments take part at EconStor, as an initial repository, and AGRIS as a target repository. Hence, the selected repositories are an integral part of the experimental setup conducted in chapters 6 and 8.

3.2 Selected Repositories

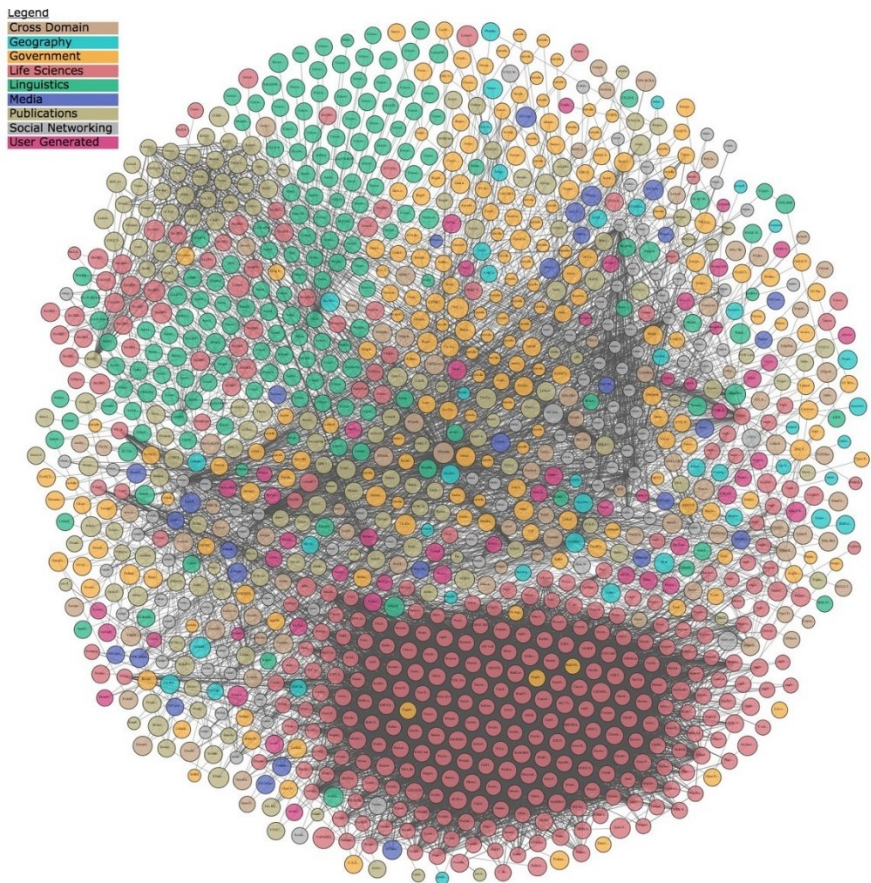


Figure 3.2 Published and interlinked datasets.¹⁰

¹⁰ Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. <http://lod-cloud.net/>

3. Semantic Web and Linked Open Data

3.2.1 EconStor Repository

In this thesis, the EconStor repository is selected as the initial repository, whose publications should be linked/enriched with information from other repositories.

EconStor is among the leading Open Access repositories in Germany and is widely related to scholarly economic literature [Oarr17]. Through EconStor, the ZBW - Leibniz Information Centre for Economics offers a platform for open access publishing to researchers in economics. It provides open access to more than 170 000 full-text documents (working papers, journal articles, conference proceedings, etc.). EconStor is used in more than 400 institutions for the digital dissemination of their publications in open access fashion. EconStor titles are visible from search engines like Google, Google Scholar, or BASE, and in academic databases like WorldCat, OpenAire, and EconBiz.

Moreover, part of EconStor metadata, i.e., 108 000 metadata records, are available as linked open data [LaBT14]. The bibliographic records are serialized as RDF triples and can be downloaded as a dump file or accessed through the SPARQL endpoint¹¹. The data are described by using vocabularies such as Dublin Core (DC), Friend of a Friend (FOAF), and RDF schema.

Table 3.1 The list of properties by vocabularies at EconStor.

Property (RDFs)	Property (FOAF)	Property (DC)
rdf:type	foaf:maker	dc:creator
rdfs:label	foaf:name	dc:description
	foaf:page	dc:issued
		dc:keyword
		dc:language
		dc:publisher
		dc:subject
		dc:title
		dc:type
		dcterms:abstract
		dcterms:isPartOf

¹¹ <http://linkeddata.econstor.eu/beta/snorql>

Table 3.1 gives the list of properties in each of these vocabularies. For description and indexing purposes, EconStor is using the Thesaurus for Economics (STW), which is also maintained by ZBW [Neub09].

3.2.2 AGRIS Repository

In achieving part of the enriching process, we are considering AGRIS as a target repository. It is one of the globally leading information systems in the area of the agricultural sciences [AJCS15]. AGRIS is a collaborative network of more than 150 institutions from 65 countries [CMWS15]. Its records, more than 7 million, are enhanced with Multilingual Agricultural Thesaurus (AGROVOC), maintained by the Food and Agriculture Organization of the United Nations (FAO) [CaKe11, CSRM12].

AGRIS is also part of the LOD cloud, by providing their data as RDF collection. At the same time, there is a SPARQL endpoint available for those interested. However, for practical reasons, we are consuming the AGRIS dump file, the version with updates of the year 2013. This dataset contains 201 038 257 statements, and for having a better usability experience, in terms of query response time and overloads, the data are stored on Ontotext GraphDB¹².

The similarity between AGRIS and EconStor is evident also in the selected RDF vocabularies. Hence, the main vocabularies used for representing the facts at AGRIS are Dublin Core (DC), BIBO, and Friend of a Friend (FOAF) [AJCS15].

3.3 Selected Thesauri

In this part, we explain two of the main thesauri used in the experimental setup regarding our evaluations. Considering the explained repositories in the previous section, the thesauri listed, i.e., STW and AGROVOC are the main indexing thesauri for the EconStor and AGRIS content.

¹² <http://graphdb.ontotext.com>

3. Semantic Web and Linked Open Data

3.3.1 STW Thesaurus

The STW Thesaurus for Economics is the leading bilingual thesaurus for economics-related content. Currently, many research institutions, development companies, and universities use it. STW has almost 6 000 subject headings in English and German and more than 20 000 additional entry terms in the economic area. The STW is developed and maintained by ZBW - Leibniz Information Centre for Economics and is continuously upgraded according to the latest changes in the economic terminology [Stw17].

The STW is also part of the Linked Open Data cloud and Semantic Web technologies [Neub09]. Through SKOS¹³ modeling scheme, STW triples are available as a downloadable file and SPARQL endpoint.

STW at the same time provides several experimental economics terminology and authority web services dedicated to humans and machines. The services primarily aim to support resource lookup and query expansion in the context of information retrieval applications. Some of these services are */suggest* for resource suggestions (starting with a given string), */synonyms* that offer alternative terms for a search term (from matching labels), */mappings* list the mappings for a concept, etc.

3.3.2 AGROVOC Thesaurus

The AGROVOC is a multilingual agricultural thesaurus, maintained by the Food and Agriculture Organization of the United Nations (FAO) [CaKe11, CSRM12]. The Thesaurus covers several areas including food, nutrition, agriculture, fisheries, forestry, and the environment. AGROVOC contains more than 32 000 concepts available in 23 languages.

AGROVOC has a wider usage by many researcher institutions, librarians and information managers for indexing, retrieving, and organizing data in agricultural information systems. It is expressed in SKOS and an LD set aligned with 16 other multilingual knowledge organization

¹³ <https://www.w3.org/TR/swbp-skos-core-spec>, accessed 12.07.2018

3.4 Aligned Concepts Between Repositories/Thesauri

systems related to agriculture. AGROVOC is downloadable as a dump file or accessible as a SPARQL endpoint.

3.3.3 WordNet Thesaurus

WordNet is a lexical database for the English language, which includes a large set of nouns, verbs, adjectives, and adverbs. One of the most interesting parts is the set of synonyms, i.e., synsets. Synonyms are meaningful related words and concepts. The WordNet content is available for download or navigable with the browser. As such, it represents a very important component for natural language processing [Mill95].

3.4 Aligned Concepts Between Repositories/Thesauri

The introduction of Linked Data concepts gives a new vision to the interoperability between different data repositories. Section 3.1 highlights more details regarding Linked Open Data, where several repositories are offering outgoing links to other repositories for interlinking the same piece of information.

Thesauri alignments represent the mappings between concepts that have the same meaning or describing the same thing. The thesauri elaborated previously, i.e., STW and AGROVOC, offer several mappings to other thesauri or vocabularies. In this way, STW thesaurus has outgoing alignments to nine other thesauri and vocabularies, according to version (v 9.08), such as Integrated Authority File (GND), DBpedia, WIKIDATA, Thesaurus Social Sciences (TheSoz), AGROVOC, German labor law thesaurus (WKD), EuroVoc, etc. Table 3.2 depicts all these mappings by also specifying the type of relations based on SKOS.

The AGROVOC thesaurus is aligned to 16 vocabularies in total, such as STW, TheSoz, DBpedia, EuroVoc, etc. Thus, there are 11 013 skos:closeMatch alignments from AGROVOC to DBpedia, 1 269 skos:exactMatch to EuroVoc, while from AGROVOC to STW there are in total 1 122 skos:exactMatch and 3 closeMatch relations.

3. Semantic Web and Linked Open Data

The benefits of such interoperability between different thesauri and vocabularies promise information retrieval operations from different repositories through the same query string. Even more, this can help overcome the language barriers, thus help the process of interlinking publications in different languages.

Table 3.2 STW mappings

AGROVOC		DBpedia	
1 027	skos:exactMatch	1 005	skos:exactMatch
1	skos:closeMatch	2 062	skos:closeMatch
German National Library (DNB)		Thesaurus Social Sciences (TheSoz)	
4 932	skos:exactMatch	3 022	skos:exactMatch
7 107	skos:narrowMatch	1 397	skos:narrowMatch
369	skos:broadMatch	81	skos:broadMatch
3 139	skos:relatedMatch	600	skos:relatedMatch
WIKIDATA		WKD German labor law thesaurus	
1 874	skos:exactMatch	270	skos:exactMatch
47	skos:closeMatch		
65	skos:narrowMatch		
16	skos:broadMatch		
8	skos:relatedMatch		

Recommender Systems and Semantic Similarity

*“Always remember that you are absolutely unique.
Just like everyone else.”*

MARGARET MEAD

4.1 Recommender Systems

The process of interlinking two items, i.e., publications or authors, is closely related to the process of recommending items based on their semantic similarity. Providing the user with the desired information, several parameters must be considered, such as a previously selected item or any other kind of preference [NMOR12]. In this way, it is inevitable to explore the application of Recommender Systems (RS) in scholarly communication, particularly in DLs [HCOC02, MoRo00, SmCa05]. The common implementation of RSs in DLs is mainly a practice used within the same repository. Therefore, recommending and interlinking publications by crosslinking relevant information from several repositories remains a challenge [DSEQ13, Hora10, Pass10a].

RSs are defined as techniques and software tools that provide suggestions for “items” to be of use to a user [RiRS11], whereas an “item” can be any piece of information that the system recommends to users. The importance of RS has been evident since the beginning of the digital era and continues to be so because of the practical application that helps users to deal with information overload [AdTu05]. Nowadays their application is in almost every field where the interaction between users and items is in focus.

4. Recommender Systems and Semantic Similarity

Through the equation 4.1, Adomavicius and Tuzhilin [AdTu05] give a more formal definition of RSs.

$$\forall c \in C, \quad s'_c = \arg \max_{s \in S} u(c, s) \quad (4.1)$$

Let c denote the users, while s the “items” recommended to the users. The utility function u that measures the usefulness of item s to user c , is defined as $u: C \times S \rightarrow R$, where R is an ordered set (e.g., nonnegative integers or real numbers within a certain range) [AdTu05]. Thus, for each user c , s' items are chosen that maximize the user’s utility. A particular rating represents the utility, which indicates the consent of that item by the user or by the system itself.

The systems for retrieving and recommending items, i.e., scientific publications, are generally grounded on content analysis, user profiles and collaborative filtering, with the incontestable role of social data as [BOHG13, LoGS11, PKCK12, SuKa10].

Hence, in this work, we follow a content analysis strategy for initiating and retrieving the list of recommended relevant resources. The approach followed here is entirely based on the set of metadata used to describe a paper in a repository, rather than any input query from the user. Thus, the extracted sets of features that characterize an item s are used for determining the similarity with the other recommended items. In essence, the user triggers the search and selects a paper from a DL that best fits her requirements. In the next step, the selected publication is enriched with closely related publications, authors, and similar information found in other repositories. The same approach is followed for recommending authors that are working on similar topics.

4.2 Similarity scoring

Determining the similarity between two texts represents a complex and challenging process. In general, there are several approaches introduced based on lexical matching, handcrafted patterns, term-weighting, and syntactic parse trees [KeRi15, RoZa10].

One of the most widely used approaches is the Vector Space Model (VSM) that represents the text documents as weighted vectors [SaMc86, SaWY75]. In this method, the TF-IDF weighting scheme is applied, while usually the similarity is measured as the cosine value between documents.

4.2.1 Vector Space Model (VSM)

The representation of a set of documents as vectors in a common vector space is known as the Vector Space Model [MaRS08, SaMc86, SaWY75]. Let's consider the set of documents $D_i = \{d_{i1}, d_{i2}, d_{i3}, \dots, d_{ij}\}$, where each of these documents is considered by terms T_j . Thus, each dimension is associated with one term, where a t -dimensional vector represents each document D_i . As mentioned before, the best-known way for calculating these dimensional values is the application of TF-IDF.

Term Frequency-Inverse Document Frequency (TF-IDF)

Typically, in content-based systems, the recommended items are text-based, such that the content is usually described with terms and keywords. However, the frequency of a term may shadow the importance of any essential term, which does not appear very often in a document. The most popular measure for determining the weights of the terms in Information Retrieval is the term frequency/inverse document frequency (TF-IDF) measure [SaBu88]. Thus, the importance of each word from the selected metadata is weighted by applying the TF-IDF algorithm [MaRS08, Ramo03].

Through the TF-IDF, each term t in document d , is weighted by a certain value, such as in the equation below (4.2), where

$$\text{TF-IDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_t \quad (4.2)$$

$\text{TF}_{t,d}$ in the basic way of interpretation, represents the number of times that a term t is into the given document d , known as local frequency. The inverse document frequency (IDF) for term t usually defined as $\log(D/n_t)$, represents the global frequency in the whole corpus for that term.

4. Recommender Systems and Semantic Similarity

Cosine Similarity (CS)

Cosine Similarity represents a standard way of measuring the similarity between two documents, d_1 and d_2 , by calculating the cosine similarity of their vector representations $\vec{V}(d_1)$ and $\vec{V}(d_2)$ [MaRS08, SaWY75]. The determination of such similarity is calculated through the equation (4.3). The numerator in this formulation represents the *dot product*, also known as the *inner product*, while the denominator represents the *Euclidean length*.

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) * \vec{V}(d_2)}{|\vec{V}(d_1)| |\vec{V}(d_2)|} \quad (4.3)$$

The implementation of averages of vectors, i.e., centroids, is a very common practice in vector space modeling. Therefore, applying CS to calculate the similarity between documents based on centroids actually represents the calculation of distance among centroids.

4.2.2 Deep Learning through Word Embedding

The lexical features, like string matching and frequency of words in a text, do not capture semantic similarity at a satisfactory level [BaDK14, KeRi15].

Current trends for determining word similarities, i.e., semantic similarities among texts, rely on vector representations of words by using neural networks, known as *word embedding* or word representations [BaDK14, BSSM06, TuRB10, CoWe08, KeRi15, KSKW15, LeCo15, LeGD15, MCCD13, MnHi09, PeSM14]. In deep learning, word embedding (WE) currently represents the most outstanding approach. Deep learning is the main discussed subject in almost every publication regarding the semantic representation of words in a low-dimensional vector [BaDK14, BSSM06, TuRB10, CoWe08, KeRi15, KSKW15, LeCo15, LeGD15, MCCD13, MnHi09, PeSM14]. Their presence is evident in many areas, such as Natural Language Processing (NLP), Information Retrieval (IR), and generation of search query strings. Word embedding inserts the complete vocabulary into a low-dimensional linear space. The embedded word vectors are trained over large collections of text corpora through neural network

models. Thus, words are embedded in a continuous vector space where semantically similar words are mapped to close vectors. Learning the word embedding is a completely unsupervised method computed on a predefined text corpus.

Word embedding currently has two well-known models of implementation: the Word2Vec algorithms proposed by Mikalov et al. for Google [MCCD13] and GloVe model from Pennington et al. at Stanford [PeSM14]. Our experiments and evaluations are based on Word2Vec due to the performance and computational cost.

Word2Vec Embedding

As noted before, Word2Vec is a novel word embedding approach, which learns a vector representation for each word using the neural network language model [MCCD13]. Two implementations of Word2Vec can be found, the continuous bag-of-words (CBOW) and skip-gram. CBOW predicts a word from the context of input text (surrounding words); while Skip-gram predicts the input words from the target context, (surrounding words are predicted from one input word). Word2Vec uses the hierarchical softmax training algorithm, which best fits for infrequent words while negative sampling is used to frequent words and low dimensional vectors. Based on the previous analyses in [BaDK14, KSKW15, MCCD13], the skip-gram model with the use of the hierarchical softmax algorithm is particularly efficient regarding the computational cost and performance. CBOW is recommended as more suitable for larger datasets. As such, the model can be trained on conventional personal machines with billions of words, achieving the ability to learn complex word relationships [KSKW15, MCCD13].

Currently, there are several implementations of Word2Vec in different environments. The native proposed code is optimized in the C programming language. However, Deeplearning4j implements a distributed form of Word2Vec for Java and Scala, while Gensim and TensorFlow offer a Python implementation of Word2Vec.

4. Recommender Systems and Semantic Similarity

4.3 Ranking evaluation metrics

One of the most notable metrics for quantifying the performance of ranking high relevant documents is Discounted Cumulative Gain (DCG) measure [JäKe00]. The formulation of DCG is defined as below, where the main inputs are the relevance value of the retrieved documents (rel_i) with the corresponding ranked positions (i); n represents the number of evaluated documents.

$$DCG_n = \sum_{i=1}^n \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^n \frac{rel_i}{\log_2(i+1)} \quad (4.4)$$

It is worth mentioning that several other modifications of DCG are present for different circumstances. Thus, for a more general representation of these values, the normalized discounted cumulative gain (nDCG) is applied. The nDCG represents the fraction of DCG with ideal DCG ($nDCG = DCG / IDCG$). Finally, ideal DCG is the recalculation of DCG after sorting the retrieved documents in decreasing order of relevance.

Part II

Linking Science

Research Design

*“there is nothing permanent
except change”*

HERACLITUS

The second part of the thesis describes the main approaches to crosslink and enriches information between publications and authors. This chapter focuses on the general strategies followed to this purpose, including the approaches to identify and make use of this data. The next two chapters, namely, 6 and 7, provide a detailed overview of the proposed approaches.

Information retrieval in Digital Libraries has been an issue since the beginning of their creation. Their application in DLs shows different approaches, such as term matching, statistical analyses of text, i.e., word frequency, or the semantic approach, by searching concepts rather than words [Scha97]. Therefore, for facilitating the process in most of the cases, publications are enriched with terms from a subject thesaurus. Similarly, IR has also been seen as an alternative for achieving interoperability among resources from different domains [LyGa96]. However, the interoperability of digital resources still continues to be one of the issues faced in the world of scholarly communication. Even today DLs are considered as isolated silos where in some instances it is almost impossible to cross the boundaries by spreading one’s search query into several sources, i.e., DLs. The lack of interoperability is even more evident when we try to handle crossdomain repositories starting from an initial DL.

The intention behind this work is to emphasize the advantages which result from an improved interoperability among different DLs and to further investigate different approaches for achieving it. With this goal in mind, we are considering including other related information that exists about the publication, i.e., other publications from other disciplines,

5. Research Design

authors' details, co-authors relations, information about the institute or organization, events, etc.

This chapter begins by highlighting the main idea and the potential sources for accomplishing the interoperability in order to enrich a DL with additional information. Continuing with the proposed research approaches, this goes through different directions before converging to the same end result. The first path is initiated from the publications metadata, while the second path has the author metadata as a starting point. Therefore, section 5.4 in detail explains the set of publication's metadata at the initial repository, including the information from the used thesaurus, by following with similar details about the publications found in the targeted repository. The chapter ends with the author's main metadata at the initial repository and VIAF clusters information.

5.1 The aim

Enriching the content of a DL with additional information from other DLs, especially regarding information that is related to a publication and author, is defined as the aim in our case. Therefore, we raise the need to add other information about a publication, such as retrieving closely related publications from other repositories and domains, or providing a wider profile of an author with a more detailed description. In this way, starting from a single point of access, a scholar would be able to find an enriched resource with additional information, rather than navigating in different places to accomplish her request.

In essence, the user triggers the search and selects a paper or an author from a DL that best fits her requirements. Thus, in the next step, the selected publication would be enriched with closely related publications, or similar information found in other repositories. In the case of the author, it can be the authors' related data such as co-authors relations and other publications similar to her field, or some non-library resources such as biographical details, affiliations, professions, etc.

5.2 Identifying sources for enrichment

In order to achieve enriched resources within a DL, it is necessary to point out relevant data from other sources, as well as to find the most appropriate approaches to crosslinking and harvest the information.



Figure 5.1 Enriching a scientific publication or author with other related information

Scenario:

“Starting from a single point of access, a scholar would be able to find resources, i.e., publications and authors, previously enriched with several other information from different repositories, that may belong to entirely different areas, but semantically similar to the initial publication. When a scholar fetches a publication in a DL, the system will offer her a list of semantically related publications from other repositories, an extended list of co-authors, and other related data corresponding to that publication.”

5.2 Identifying sources for enrichment

To fulfill the aim of enriching the content of a publication or author within a DL, primarily we need to identify the sources from which the data can be retrieved. For this purpose, several paths will be followed. As one of the possible sources for data enrichment, we consider bibliographic and authority repositories which are offered by several libraries and institutions. The main direction will be to leverage the already available content on the semantic web, such as Linked Open Data (LOD) repositories,

5. Research Design

as one of the most promising data sources [BHIB08, FMFG18]. In this way, repositories available as semantic web content, such as bibliographic Linked Open Data (LOD) repositories are in the focus of this study. As described in chapter 3, we firstly consider the existing alignments among concepts between repositories and exploring best practices for consuming them in the context of crosslinking information. The initial experiments are done between EconStor and AGRIS repositories, based on structural similarity between them (i.e., vocabularies and thesauri) and the crossdomain background (i.e., economics vs. agriculture). Both of them offer an open catalog as part of LOD cloud with available SPARQL endpoints and RDF dump files, as well as thesauri named STW and AGROVOC, respectively. In section, 3.2.1 and 3.2.2 provide more details regarding these repositories. After that, we investigate the role of thesauri, including descriptors with the corresponding narrowed, broadened, and extended concepts through a Simple Knowledge Organization System Reference - SKOS¹⁴ vocabulary. For this purpose, as described in sections 3.3.1 and 3.3.2, the STW and AGROVOC thesauri are explored.

Additionally, any external service such as WordNet (section 3.3.3), would be considered for extending the list of terms and concepts. Furthermore, several DLs are offering API services for accessing their catalogs or any other particular information, such as the title or abstract of a publication. For instance, such a service is provided by the German National Library (DNB) that makes it possible to explore the catalog and extract the record extraction in different representations.

In the intention to crosslink and enrich author-related information, the approaches for authority file aggregations are among the first we explore. Several approaches that uniquely identify and produce correlations between researchers [PaKS15] are also considered. In section 2.3 we provide a list of the most prominent services. The purpose of their usage is related to author name disambiguation for achieving crosslinking information among different repositories. For instance, VIAF and WIKIDATA are considered as the most relevant services in our scenarios. Sections 2.3.1 and 2.3.2 provide more details about the two.

¹⁴ <https://www.w3.org/TR/swbp-skos-core-spec>, accessed 12.07.2018

5.3 Proposed approaches

In section 5.1 the main goal of the thesis is about enriching/crosslinking a particular resource, be it a publication, or an author, inside a DL is already elaborated. Therefore, for each possible resource, a wider profile will be generated and enhanced with additional information. Based on this, the main challenge is about achieving the right approach for identifying, using, and evaluating the targeted information. The process of crosslinking data from different repositories is crucial for this goal.

The approach followed in this work is entirely based on the set of metadata that are used to describe a paper in a repository, rather than any input query from the user. For this purpose, we have followed two main approaches for achieving interoperability and retrieving the most relevant information from the targeted repositories. Actually, in both cases, the interoperability is initiated from one repository, i.e. a particular DL, in which resources are intended to be enriched with additional information. Thus, in this document, we will refer to it as the “initial repository”, while repositories where we try to find and retrieve the data as “target repositories”. Appendix A.1 provides detailed descriptions of these definitions.

- The first approach is related to **publication-centered metadata**. Primarily, the existing alignments among the concepts between LOD repositories (thesauri) will be considered, by exploring best practices for consuming them. Improvements regarding the semantic measurements between resources are achieved by evaluating several text-mining and machine learning techniques. Chapter 6 gives the details of this approach.
- The second approach is related to **author-centered metadata**. Therefore, for a given author, we find the correlations with other authors, publications or other related information by crosslinking data. Additionally, before the author profile enrichment, the process of author name disambiguation and accurate identification is applied, as a mandatory step to harvest and crosslink information. This approach is presented in chapter 7.

5. Research Design

5.4 Publication-centered metadata at initial repository

The initial experiments regarding the interoperability are done between EconStor and AGRIS, based on the structural similarity between these two repositories, such as metadata of the collections that they host, used vocabularies, as well as the presence of thesauri on both sides, STW and AGROVOC respectively. The reason we choose these two repositories is that they support our goal for interlinking repositories from different disciplines. Both repositories offer an open catalog, part of LOD cloud, with available SPARQL endpoints and RDF dump files.

Analysis of the existing metadata, which are used to describe a publication in the initial repository, i.e., EconStor will be one of the first steps to achieve the interoperability goal. A wide range of metadata describes each paper in EconStor. Besides the common ones for title, abstract, authors, year and publisher, the application of the STW thesaurus provides enrichment with a huge set of descriptors and concepts with the respective mappings to other repositories.

As mentioned in section 3.2.1, part of the EconStor records is available as LOD. Thus, through SPARQL queries we are retrieving the necessary information concerning a publication. The SPARQL query listing 5.1 shows such an example, for having the title, abstract, hyperlink, year and publisher.

Listing 5.1 Retrieving publication's information from EconStor SPARQL endpoint

```
SELECT DISTINCT ?p ?title ?abs ?hlink ?issued ?publ
WHERE {
    ?p    dc:title ?title;
         foaf:page ?hlink;
         dc:publisher ?publ.
    OPTIONAL {?p dcterms:abstract ?abs}.
    OPTIONAL {?p dc:issued ?issued}.
}
```

Moreover, though the listing 5.2 we able to retrieve the list of authors that are assigned to that publication. Therefore, there are some instances when more than one SPARQL query requests need to be executed to obtain the necessary information.

5.4 Publication-centered metadata at initial repository

Listing 5.2 Retrieving publication's author(s) from EconStor SPARQL endpoint

```
SELECT DISTINCT ?name
WHERE {
    ?p  dc:creator ?a.
    ?a  foaf:name ?aname.
    OPTIONAL {?p dc:creator ?other.
               ?other foaf:name ?name}.
}
```

In order to get an overall view of the dataset collections, we have built a prototype that retrieves and presents this information. Through it, we can display and group information in a comprehensive way, but also perform a series of experiments using this data. Searching for a particular EconStor publication, the user is provided with data as presented in figure 5.2. Each of the constituting elements of that result set are denoted in the following way: title (p_t), abstract (p_{abs}), hyperlink (p_h), year (p_y), publisher (p_p), keywords (K_p) and synonyms for the showed keywords (S^k). It is worth mentioning that the synonyms are generated by consuming the existing STW web service, *econ-ws/synonymss*¹⁵, that returns a set of alternative terms for a given term. Furthermore, in many cases, we have considered the usage of WordNet synonyms as part of our experimental setups.

The use of STW thesaurus obviously affects the description of publications by enriching the metadata set with several descriptors. The SKOS modeling scheme, i.e. vocabulary, supports describing resources with several descriptors, which over the same scheme are narrowed, broadened or presented with related terms. Besides these, the STW thesaurus provides alignments among concepts between repositories. Thus, each concept found behind an EconStor publication is mapped to a concept with the same meaning in other repositories (see section 3.4). Figure 5.3 gives a closer view of one of these descriptions used for describing a paper, i.e. *Inflation*. From here, based on STW thesaurus, *inflation* narrowed to *Stagflation*, *Hyperinflation* and *Core inflation*, broadened to *Price level*, while related to *Anti-inflation policy*, *Inflation theory*, *Inflation rate* and *Wage-price spiral*.

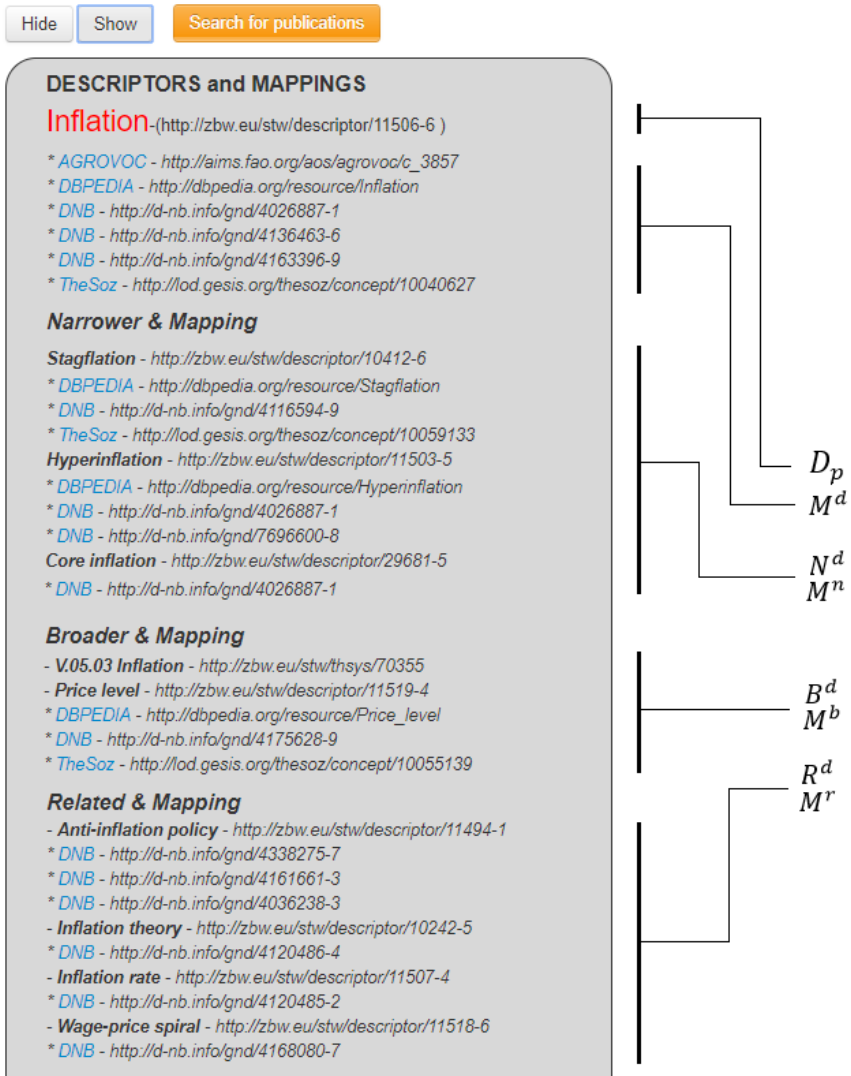
¹⁵ <http://zbw.eu/beta/econ-ws/about>, accessed 13.06.2018

ECONSTOR

PROTOTYPE

Figure 5.3 also makes visible the outgoing links of concept to other linked open data repositories and vocabularies. As can be noted, the concept “Inflation” through STW is aligned to DBpedia, AGROVOC, German National Library (DNB) and TheSoz. However, except “Inflation”, the chosen publication has five more descriptors, such as *Corporate taxation*, *Equity capital*, *Bank*, *Banking history* and *Sweden*, which are also mapped to similar vocabularies. Figure 5.4 gives a summary of these descriptors, where in total four of them are aligned to AGROVOC, by including the corresponding links.

5.4 Publication-centered metadata at initial repository



5. Research Design

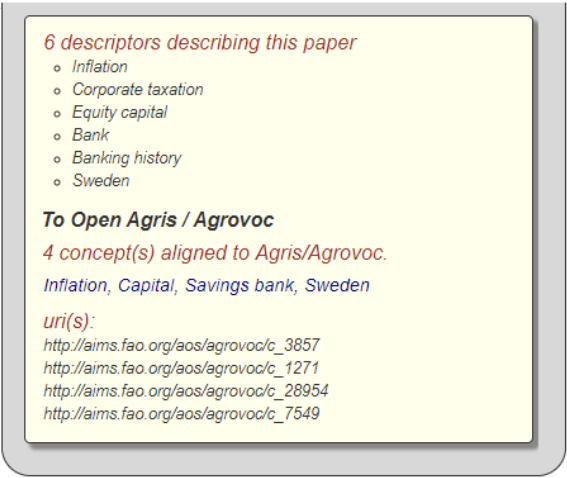


Figure 5.4 Summary of main descriptors by highlighting AGRIS alignments.

The table below represents the notation of the selected metadata from the initial repository, which are further used in EconStor experiment.

Table 5.1 The notation table – publication’s metadata from the initial repository.

Notation	Description
p_t	title
p_{abs}	abstract
p_h	publication’s hyperlink at EconStor
p_y	published year
p_p	publisher
$A_p = \{a_1^p, a_2^p, a_3^p, \dots, a_n^p\}$	authors of publication p
$K_p = \{k_1^p, k_2^p, k_3^p, \dots, k_n^p\}$	<i>dc:keyword</i> of publication p
$S^k = \{s_1^{k1}, s_2^{k1}, \dots, s_1^{k2}, s_2^{k2}, \dots, s_{kn}^{pn}\}$	synonyms of the keywords k
$D_p = \{d_1^p, d_2^p, d_3^p, \dots, d_n^p\}$	<i>dc:subject</i> , main descriptors of publication p
$N^d = \{n_1^{d1}, n_2^{d1}, \dots, n_1^{d2}, n_2^{d2}, \dots, n_{kn}^{dn}\}$	<i>skos:narrower</i> concepts for the descriptor d
$B^d = \{b_1^{d1}, b_2^{d1}, \dots, b_1^{d2}, b_2^{d2}, \dots, b_{kn}^{dn}\}$	<i>skos:broader</i> concepts for the descriptor d
$R^d = \{r_1^{d1}, r_2^{d1}, \dots, r_1^{d2}, r_2^{d2}, \dots, r_{kn}^{dn}\}$	<i>skos:related</i> concepts for the descriptor d
M^d, M^n, M^b, M^r	mappings for aligned concepts

5.4 Publication-centered metadata at initial repository

5.4.1 Publication-centered metadata at target repository

As mentioned previously, the initial experiments in this thesis are done by crosslinking information between two LOD repositories, i.e., EconStor and AGRIS. Therefore, as a target repository at this phase, we point to AGRIS, which serves as the multilingual bibliographic database for agricultural science and technology. More detailed information about this repository is shown in section 3.2.2, including the complementary AGROVOC thesaurus in section 3.3.2.

In general, we are focused on the same set of metadata as in our initial repository. Hence, for each publication d , from AGRIS, we consider the title (d_t), abstract (d_{abs}) and other general data; keywords in this case are not provided. Moreover, using the AGROVOC thesaurus, the metadata set is extended with the main descriptors (D_d), including the narrowed, broadened, and related terms, similarly as in our initial repository.

Regarding AGRIS, the version with updates of the year 2013 is loaded, with 201 038 257 RDF statements. The datasets of AGRIS and AGROVOC are stored locally using the Ontotext GraphDB data storage component. The data are consumed by executing several SPARQL queries. The listing 5.3 shows the example for retrieving the main AGRIS metadata including the links of descriptors.

Listing 5.3 Retrieving publication's metadata including the main descriptors from the target repository (AGRIS).

```
SELECT ?d ?title ?abs ?uri ?year (GROUP_CONCAT( ?subject; SEPARATOR = ",") AS ?desc)
{
  SELECT distinct ?d ?subject ?title ?abs ?uri ?year
  WHERE{
    ?d ?p ?o;
    dcterms:subject ?subject;
    dcterms:title ?title;
    bibo:abstract ?abs;
    dcterms:language "eng".
    OPTIONAL{?d bibo:uri ?uri.}
    OPTIONAL{?d dcterms:issued ?year.}
  }
}
GROUP BY ?d ?title ?abs ?uri ?year
```

5. Research Design

Table 5.2 A sample of retrieved metadata for an AGRIS publication.

Id	Title	Abstract	Link	Year	Descriptors
http://agris.fao.org/aos/records/US201301364477	Are government regulations pushing food prices higher?	Abstract: Several "consumer protection" bills before Congress may serve to raise food prices, and a report from GAO tries to explain the complex food-price situation and what government and food industry can do to help. Causes of food price rises are described, and three recommendations to the food industry to improve efficiency and lower costs are given: computerized check-out systems, minimizing food loss, and standardizing containerization (modularization). The GAO emphasizes that government decision makers need to consider effects of their actions on food industry costs and related food prices. Increased costs to consumers should balance expected benefits. A chart showing proposed government actions that could effect food costs is included.	NA	1978	http://aims.fao.org/aos/agrovoc/c_3020 , http://aims.fao.org/aos/agrovoc/c_3857 , http://aims.fao.org/aos/agrovoc/c_1358166351183

As an instance of the query output from the listing 5.3 generates a view as shown in table 5.2. Based on this metadata set, we are considering the title and abstract of crucial importance, especially in the steps when different data mining approaches are applied. In addition, the assigned descriptors represent a valuable input in this regard, based on the fact that they derive from a controlled vocabulary and annotated under the care of domain experts. The example in table 5.2 shows three descriptors with the link to the corresponding term, including the id at the end (e.g. c_3020). For retrieving the list of labels, instead of hyperlinks, we refer to the AGROVOC thesaurus. The SPARQL query given below listing 5.4 provides an example where the output of the chosen concept will consist of a single label, which in this case it is “Inflation”.

Listing 5.4 Retrieving descriptor’s label from AGROVOC

```
SELECT DISTINCT ?mainLabel
{
  <http://aims.fao.org/aos/agrovoc/c\_3857> skos:prefLabel ?mainLabel.
  FILTER (langMatches(lang(?mainLabel), "EN"))
}
```

5.4 Publication-centered metadata at initial repository

In this case, query results are limited only to the English language, avoiding the results that contain multilingual labels from AGROVOC. Otherwise, the label of the concept “Inflation” is provided in 22 languages, which also can be a powerful point for achieving cross language interoperability. However, due to the multilingual noise in the initial repository, and for the sake of evaluations later on in the research, we have focused on English language publications.

Staying at the same concept (Inflation), its description can be extended by listing the narrowed, broadened or related terms. The example in listing 5.5 shows exactly such an instance, in which case the thesaurus provides only one additional term as a broadened concept, i.e., “monetary policies”. All the terms related to a given concept, represent an important component for further steps when data mining approaches are considered.

Listing 5.5 Extracting narrowed, broadened and related concepts of a selected descriptor from AGROVOC thesaurus.

```
SELECT DISTINCT ?label
WHERE
{
  {
    <http://aims.fao.org/aos/agrovoc/c\_3857> skos:narrower ?nar.
    ?nar skos:prefLabel ?label.
  }
  UNION { <http://aims.fao.org/aos/agrovoc/c\_3857> skos:broader ?br.
    ?br skos:prefLabel ?label.
  }
  UNION { <http://aims.fao.org/aos/agrovoc/c\_3857> skos:related ?re.
    ?re skos:prefLabel ?label.
  }
  FILTER (langMatches(lang(?label), "EN"))
}
```

It is worth mentioning that besides the fact that the experiments are based on one repository such as AGRIS, the proposed approaches can be evaluated without any additional effort at any repository, and this is so especially in case of providing a similar set of metadata. Even in the case when only the title or abstract is available, one of the proposed methods such as word embedding performs at a satisfactory level, (section 6.3).

5.5 Author-centered metadata at initial repository

In the second approach, the data crosslinking process relies on the metadata set that is used to describe an author in a particular repository. Therefore, the procedure starts from an initial repository, i.e., EconStor, whose authors to be uniquely identified and enriched with other data. The most basic metadata for describing an author are Name and Surname. Hence, each 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 the 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, as we have presented earlier [HaRT15], 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 $A_a = \{a_1^a, a_2^a, a_3^a, \dots, a_n^a\}$. 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, $\bar{P}_a = \{\bar{p}_j^{ai}; i = 1, n; j = 1, k\}$.

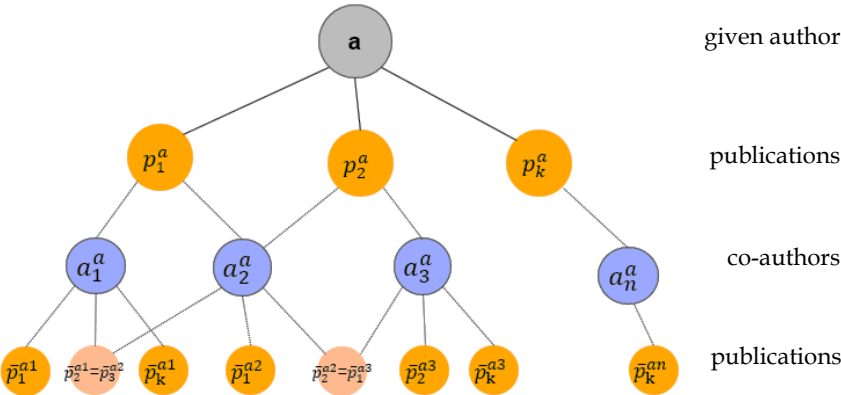


Figure 5.5 The relationship among authors, co-authors, publications and co-authors publications for a given author \mathbf{a}

5.5 Author-centered metadata at initial repository

Table 5.3 represents the set of these metadata. A detailed picture of the relationships is shown in Figure 5.5, where can be seen that p_1^a and p_2^a have a common author.

Table 5.3 Notation table – author’s metadata from the initial repository

Notation	Description
$\mathbf{a}, \mathbf{a} = (t_1, t_2).$	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
$A_a = \{a_1^a, a_2^a, a_3^a, \dots, a_n^a\}$	co-authors of the author \mathbf{a}
$\bar{P}_a = \{\bar{p}_1^{a1}, \dots, \bar{p}_k^{a1}, \dots, \bar{p}_1^{a3}, \dots, \bar{p}_k^{an}\}$	publications of co-authors of \mathbf{a}

5.5.1 VIAF metadata

We are considering VIAF clusters as “bridges” for achieving the disambiguation and crosslinking process concerning the author’s related approach. Therefore, several operations are taken 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 $j=1, k$ (k is the number of retrieved clusters that can be different in individual cases).

Inside each of these VIAF clusters, different forms of authors’ name alternatives can be found for a particular author, obtained from the native libraries, as shown in figure 5.6. Henceforth, the set of variations is denoted as $A_{c_j} = \{a_1^{cj}, a_2^{cj}, a_3^{cj}, \dots, a_l^{cj}\}$, where each certain name alternative is given such as $a_1^{cj} = (t_1, t_2)$, similarly as in the initial repository. Except for 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. Given that the set of publications and co-authors is an important piece of information in assessing the cluster's importance to a particular author, we are denoting them as in the following. The set of publications found in a particular cluster is notated with $P_{c_j} = \{p_1^{cj}, p_2^{cj}, p_3^{cj}, \dots, p_k^{cj}\}$, while the set of co-authors within a cluster will be $\hat{A}_{c_j} = \{\hat{a}_1^{cj}, \hat{a}_2^{cj}, \hat{a}_3^{cj}, \dots, \hat{a}_n^{cj}\}$.

5. Research Design

Besides these data, the set of publications retrieved directly from the libraries or institutions that are contributing to that cluster can be of particular importance. These publications can be retrieved by referring to 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 $\check{P}_{cj} = \{\check{p}_1^{cj}, \check{p}_2^{cj}, \check{p}_3^{cj}, \dots, \check{p}_k^{cj}\}$.

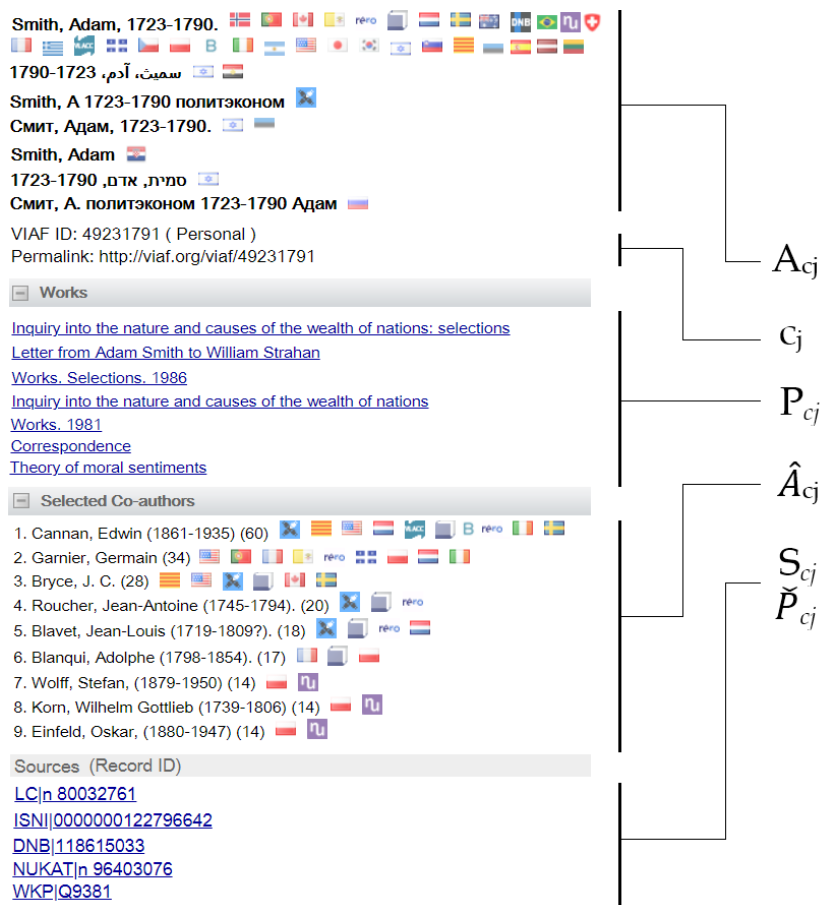


Figure 5.6 A particular cluster (heading) in VIAF

5.5 Author-centered metadata at initial repository

In conclusion, the ultimate set of metadata from a particular VIAF cluster that we are considering are the most important in our experimental setup is presented in table 5.4.

Table 5.4 Notation table – author’s metadata from a VIAF cluster

Notation	Description
C_j	clusters to be checked at VIAF
$A_{c_j} = \{a_1^{c_j}, a_2^{c_j}, a_3^{c_j}, \dots, a_l^{c_j}\}$	author’s names variations in a VIAF cluster c_j , $j=1, k$
C_{byear}	birth year
C_{dyear}	death year
$P_{c_j} = \{p_1^{c_j}, p_2^{c_j}, p_3^{c_j}, \dots, p_k^{c_j}\}$	publications in a VIAF cluster c_j
$\hat{A}_{c_j} = \{\hat{a}_1^{c_j}, \hat{a}_2^{c_j}, \hat{a}_3^{c_j}, \dots, \hat{a}_n^{c_j}\}$	co-authors in a VIAF cluster c_j
$S_{c_j} = \{s_1^{c_j}, s_2^{c_j}, s_3^{c_j}, \dots, s_k^{c_j}\}$	sources to other places from a VIAF cluster
$\check{P}_{c_j} = \{\check{p}_1^{c_j}, \check{p}_2^{c_j}, \check{p}_3^{c_j}, \dots, \check{p}_k^{c_j}\}$	publications from other sources in the VIAF cluster

Linking publications across different LOD repositories

*“What we know is a drop,
what we don't know is an ocean.”*

Isaac Newton

As noted in chapter 5, the process of crosslinking information from different repositories represents the key step to the ultimate goal, the enrichment of DL resources with additional information. In our approach, the interoperability is initiated from one repository i.e., DL, by considering all existing metadata for a single publication. Using this information, we are connecting to other external repositories to search for possible semantically related publications and other related information (e.g. author details) to the initial publication. In order to achieve this, we leverage already available content on the semantic web, such as Linked Open Data (LOD) repositories, as one of the most promising data sources [BHIB08, BHLO01, LaST16]. As such, the existing alignments among concepts between repositories are considered with the corresponding narrowed, broadened and extended concepts through the SKOS modeling scheme. Parts of this chapter are published at several proceedings and journals [HaLT14, HaTo16, HaTo17, HaTo18].

For retrieving a set of publications as semantically similar to the initial publication, the application of semantic technologies, information retrieval and machine learning methods are applied. For this purpose, we present preliminary experiments conducted by Vector Space Models (VSM) [SaWY75] through the application of TF-IDF and Cosine Similarity (CS). Special attention is given to the process of determining key concepts at the initial publication, as an essential point to initiate the crosslinking process.

6. Linking publications across different LOD repositories

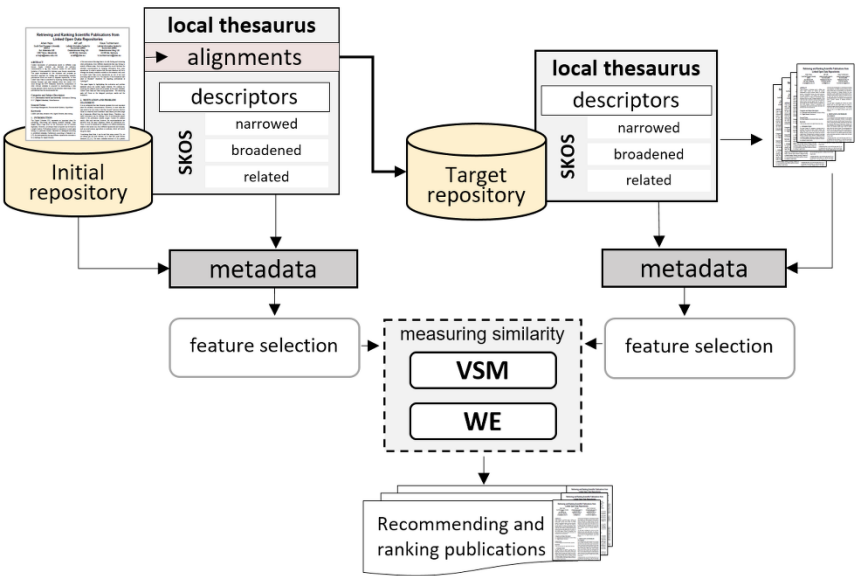


Figure 6.1 Enriching a scientific publication with recommendations from LOD repositories.

Additionally, we extend the experiments by applying a Word Embedding (WE) approach, in which we are focusing mainly on the context of distributed word representations, instead of words frequency, weighting and string matching. The contemporary Word2Vec implementation is applied as a similar Deep Learning approach to model semantic word representations [MCCD13]. An ultimate overview of this approach is represented in figure 6.1.

6.1 Using the aligned concepts

Currently, a large number of libraries have exposed their data as RDF statements inside the LOD cloud. Such example are German National Library (DNB), Library of Congress (LoC), Swedish National Library (LIBRIS), British National Bibliography (BNB), Europeana Digital Library,

6.1 Using the aligned concepts

Leibniz Information Centre for Economics (ZBW), Food and Agriculture Organization of the United Nations (FAO), DBLP Bibliography Database, etc. Most of these LOD repositories as part of LOD cloud, offer a number of incoming/outgoing links to other repositories for mapping several resources or concepts that have the same meaning.

As mentioned in section 3.4, EconStor through the STW thesaurus has numerous mappings to other thesauri and vocabularies. For instance, for AGROVOC 1 027 *skos:exactMatch* alignments exist while to DBpedia 1 005 *skos:exactMatch*. Therefore, the interlinking process primarily is initiated from the existing alignments among concepts between repositories, by exploring best practices for consuming these mappings.

Let us elaborate this with some examples. The STW concept “Biofuel” is used for describing and indexing several publications in the EconStor repository. Hence, a concept with similar meaning may exist in several other thesauri, for describing publications in their initial repositories. In particular, the same concept is present at AGROVOC thesaurus with the label “biofuels”, used for indexing publications at AGRIS repository. The interlinked concepts in figure 6.2 give a better interpretation of this idea.

Therefore, the alignment between these two concepts would make it possible to retrieve all the publications from both repositories with the same query, using “biofuel”. Thus, starting from EconStor repository, by using that concept we are able to retrieve all the publications from AGRIS that having exactly that concept among the selected descriptors. Hence, a simple query shows that the concept “biofuel” is used for describing 7 083 documents in AGRIS catalog. However, since a particular publication may be described by numerous descriptors, exists the possibility of several of them to be aligned. Thus, the number of recommended publications from other repositories, based on these alignments is on different size. Through the SKOS schema, hierarchical conceptual navigation can be performed in the initial or at the target repository. All this has an impact on the selected concepts by narrowing and broadening the set of results.

In addition, the same concept can be aligned to other repositories at the same time. Thus, if for example, AGROVOC provides outgoing alignment for the same concept to another repository, which alignments are missing in the initial repository, the request can be distributed there too. Figure 6.2 gives a visual understanding of this indication.

6. Linking publications across different LOD repositories

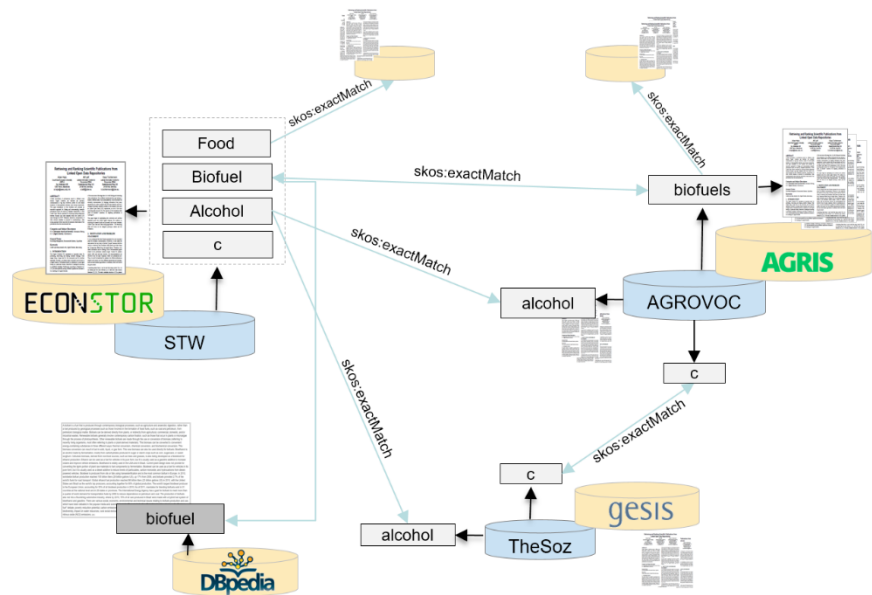


Figure 6.2 Thesauri alignments

After a closer view of these components and a variety of experiments, there are notable two phenomena. Firstly, there is obvious that not all the publications inside a repository are described with descriptors. Inside the EconStor repository cases without descriptors are very rare, however, there are several cases where descriptors are so general terms, such as “Theory” or “Germany”. Secondly, not all descriptors are aligned with concepts from other repositories. Therefore, the experiments are focused on publications that have at least one descriptor with the outgoing link to the targeted repository.

By triggering an EconStor publication, the developed prototype makes it possible to show all the available metadata behind that publication. Besides the common metadata, explained in section 5.4, all the alignments to other repositories and thesauri are highlighted with the pointing repositories in particular. Figure 6.3 gives an example of considering the descriptor “Inflation”.

6.1 Using the aligned concepts

Inflation-(<http://zbw.eu/stw/descriptor/11506-6>)

* **AGROVOC** - http://aims.fao.org/aos/agrovoc/c_3857

* **DBPEDIA** - <http://dbpedia.org/resource/Inflation>

* **DNB** - <http://d-nb.info/gnd/4026887-1>

* **DNB** - <http://d-nb.info/gnd/4136463-6>

* **DNB** - <http://d-nb.info/gnd/4163396-9>

* **TheSoz** - <http://lod.gesis.org/thesoz/concept/10040627>

Figure 6.3 Alignments to other repositories for a particular STW descriptor.

Hence, the descriptor “*Inflation*” is aligned to AGROVOC, DBpedia, German National Library and TheSoz thesaurus. By analyzing the URI http://aims.fao.org/aos/agrovoc/c_3857 that points to AGROVOC, it can be seen that the concept is mapped to absolutely the same label “**inflation**”. However, this does not have always to be so; sometimes the mapped concepts can have different morphology, such as singular vs. plural (ex. Biofuel to biofuels) or a completely different label. The mapping between concepts makes it possible the terms labeled differently in separate vocabularies, to signify the same concept.

Let consider a concrete publication from EconStor title “*Do inflation and high taxes increase bank leverage?*”. In this case, in total six descriptors (D) are used for describing this paper, such as *Inflation*, *Corporate taxation*, *Equity capital*, *Bank*, *Banking history* and *Sweden*. From that list, as shown in table 6.1, only two main descriptors (D) are mapped to AGROVOC, i.e. AGRIS, by excluding the narrowed (N) and broadened (B) terms.

Table 6.1 A sample of mapped descriptors to another repository, i.e. AGRIS

	STW concept	Mapped link	AGROVOC concept
D	inflation	http://aims.fao.org/aos/agrovoc/c_3857	Inflation
D	Corporate taxation		
D	Equity capital		
B	Capital	http://aims.fao.org/aos/agrovoc/c_1271	capital
D	Bank		
N	Savings bank	http://aims.fao.org/aos/agrovoc/c_28954	savings bank
D	Banking history		
D	Sweden	http://aims.fao.org/aos/agrovoc/c_7549	Sweden

6. Linking publications across different LOD repositories

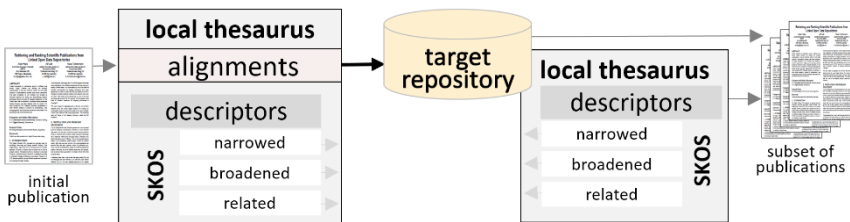


Figure 6.4 Retrieving scientific publications from LOD repositories based on concepts' alignments

In addition, the presence and the role of thesauri used for indexing the data inside repositories are investigated with particular attention. In addition to the alignments, we include the descriptors with the corresponding narrowed, broadened and extended concepts through SKOS modeling scheme. Table 6.1 gives details regarding the alignments of these concepts, where the concept "*Capital*" is denoted as broadened (B) concept from the descriptor "*Equity capital*". While the "*Bank Savings*" is narrowed from "*Bank*". The presence of such alignments can ensure a list of publications from other repositories. The idea is to retrieve publications, which are described by any of these descriptors in the target repository. Figure 6.4 shows an overview of this process.

Listing 6.1 represents a SPARQL query for getting the list of publications from a target repository where any of the mentioned descriptors are used. The example below will retrieve the publication from AGRIS that among other descriptors have the concept "*inflation*".

Listing 6.1 Retrieving publications from the target repository (AGRIS), based on a particular descriptor.

```
SELECT distinct ?d ?title ?abs ?uri ?year
WHERE{
  ?d ?p ?o;
    dcterms:title ?title;
    bibo:abstract ?abs;
    dcterms:language "eng".
  OPTIONAL{?d bibo:uri ?uri.}
  OPTIONAL{?d dcterms:issued ?year.}
  FILTER (?o=<http://aims.fao.org/aos/agrovoc/c\_7549>).
}
```

6.1 Using the aligned concepts

However, the execution of this query will result in an extremely huge list of results. In order to deliver more details, the concept “inflation” is used for describing 2 754 documents in AGRIS catalog, while “income” in 21 838. Since a publication can have several such aligned concepts, the insertion of all of them through a union is resulting in even a broader outcome. For example, to look for publications described by any of the listed concepts, the following condition might be applied:

```
FILTER (?o=<http://aims.fao.org/aos/agrovoc/c\_7549> ||  
?o=<http://aims.fao.org/aos/agrovoc/c\_3820> || ?o=<...>)
```

Meanwhile, the attempt to find publications in the target repository, with the same set of descriptors as in the initial one, results in an empty set. For example, searching for publications outlined with all the required concepts, the statement would be as follows:

```
FILTER (?o=<http://aims.fao.org/aos/agrovoc/c\_7549> &&  
?o=<http://aims.fao.org/aos/agrovoc/c\_3820> && ?o=<...>)
```

The hierarchical navigation between concepts with the use of knowledge organization systems by broadening and narrowing the concepts, e.g., the notion of Germany broadened to Europe and narrowed to Berlin, helps to reduce complexity by narrowing down the number of results. However, the choice is very arbitrary and the outcome is not satisfactory for offering a shorter list of recommended publications with the opportunity to be ranked.

Therefore, we use alignments between repositories or thesauri for retrieving an initial set of publications, especially for reformulating a search query from one vocabulary to another [BiTu16, HaLT14, JJHY12]. The importance of these descriptors, as well as the alignments among them, is considered as undisputed since experts in relevant fields set them manually.

The presence of thesauri in the targeting repository can be useful for extending the corpus of metadata concepts, which, as we will show later, is very significant for further analyses. Thus, besides the hierarchical navigation at the initial repository, through the alignments, it is possible to perform the same steps in the target repository. Such an example provides the case in table 6.1. Accordingly, the concept “inflation” is broadened to

6. Linking publications across different LOD repositories

“monetary policies”, *“capital”* narrowed to *“fixed capital”* and *“working capital”*, *“savings bank”* is broadened to *“banks”* while *“Sweden”* is broadened to *“Scandinavia”*. All these newly founded concepts can be part of the publication metadata at the initial repository, for further text mining steps. Apart from this information, almost each of these labels is provided in more than 16 languages. In several cases, the target repository offers a short definition of the concepts, thus for example, *“inflation”* is defined at AGROVOC as *“the overall general upward price movement of goods and services in an economy”*. Moreover, the target repository also provides outgoing links to other repositories for the selected concept. This is a good opportunity to extend the mapping list of that concept to the initial repository with the newfound outgoing links.

The *“inflation”* at AGROVOC is mapped to five other repositories, from which three of them are not in the STW mapping list for *“inflation”*.

- CAT (http://cat.aii.caas.cn/concept/c_45316), not in the initial repository
- DNB (<http://d-nb.info/gnd/4026887-1>)
- Eurovoc, (<http://eurovoc.europa.eu/1421>), not in the initial repository
- USDA (<http://lod.nal.usda.gov/nalt/28678>), not in the initial repository
- ZBW (<http://zbw.eu/stw/descriptor/11506-6>)

However, in addition to AGROVOC links, STW thesaurus offers a mapping to several other repositories. Section 3.4 gives detailed information about these alignments. Therefore, the presence of DBpedia mappings can be used for having some general information about a particular concept. Hence, by performing a query as in the listing 6.2, the abstract and the Wikipedia link are retrieved from the DBpedia repository. Actually, the prototype provided a visual view of such output for every concept behind an EconStor publication (STW concept) if the DBpedia mapping is provided. Figure 6.5 gives an overview of such an instance where the definition of *“Inflation”* is generated.

Listing 6.2 Getting DBpedia information about an STW concept.

```
SELECT distinct ?abs ?link WHERE {  
<http://dbpedia.org/resource/Inflation> dbo:abstract ?abs;  
foaf:isPrimaryTopicOf ?link .  
FILTER (langMatches(lang(?abs), "en")).  
}
```

6.1 Using the aligned concepts

Abstract:

In economics, inflation is a sustained increase in the general price level of goods and services in an economy over a period of time. When the price level rises, each unit of currency buys fewer goods and services. Consequently, inflation reflects a reduction in the purchasing power per unit of money – a loss of real value in the medium of exchange and unit of account within the economy. A chief measure of price inflation is the inflation rate, the annualized percentage change in a general price index, usually the consumer price index, over time. The opposite of inflation is deflation. Inflation affects economies in various positive and negative ways. The negative effects of inflation include an increase in the opportunity cost of holding money, uncertainty over future inflation which may discourage investment and savings, and if inflation were rapid enough, shortages of goods as consumers begin hoarding out of concern that prices will increase in the future. Positive effects include reducing the real burden of public and private debt, keeping nominal interest rates above zero so that central banks can adjust interest rates to stabilize the economy, and reducing unemployment due to nominal wage rigidity. Economists generally believe that high rates of inflation and hyperinflation are caused by an excessive growth of the money supply. However, money supply growth does not necessarily cause inflation. Some economists maintain that under the conditions of a liquidity trap, large monetary injections are like "pushing on a string". Views on which factors determine low to moderate rates of inflation are more varied. Low or moderate inflation may be attributed to fluctuations in real demand for goods and services, or changes in available supplies such as during scarcities. However, the consensus view is that a long sustained period of inflation is caused by money supply growing faster than the rate of economic growth. Today, most economists favor a low and steady rate of inflation. Low (as opposed to zero or negative) inflation reduces the severity of economic recessions by enabling the labor market to adjust more quickly in a downturn, and reduces the risk that a liquidity trap prevents monetary policy from stabilizing the economy. The task of keeping the rate of inflation low and stable is usually given to monetary authorities. Generally, these monetary authorities are the central banks that control monetary policy through the setting of interest rates, through open market operations, and through the setting of banking reserve requirements.

WIKIPEDIA:
[WIKIPEDIA Link](#)


Publications:
1.  http://www.nobelprize.org/nobel_prizes/economics/laureates/1976/friedman-lecture.pdf

Figure 6.5 DBpedia information for a selected STW concept, retrieved through the prototype

The performed experiments show that the alignments among concepts are an important element to break the heterogeneity between vocabularies and crosslink resources from different repositories. However, for the reasons outlined above, such as a wider or empty set of results, retrieving semantically similar publications based only on alignments is almost impossible. In a situation when the usage of aligned concepts generates a wider range of results, we need further processing to narrow this subset and generate a relevance-based ranking. For this purpose, the involvement of other metadata, such as title, abstract and keywords is more than required. Moreover, the presence of the thesauri descriptors affects the enrichment of this set of metadata with several other concepts, and through

6. Linking publications across different LOD repositories

the SKOS modeling scheme that set can be extended with several related terms. Section 5.4 and section 5.4.1 give details about these sets at the initial and target repository, regarding our experimental setups.

Therefore, by holding all these metadata elements, the implementation of data mining approaches is considered. In two different approaches, we try to measure the semantic similarity between the triggered publication at the initial repository, with the retrieved publications as result of alignments from the targeted repositories. The process begins with one of the most essential and widely used approaches for this purpose, such as the Vector Space Model.

6.2 Vector Space Model approach

The terms extracted from the publications metadata, at the initial and target repositories are represented as separate vectors through the Vector Space Model. Therefore, the terms of a publication from the initial repository are projected in the vector $\vec{V}(p)$, while each publication in the target repository will embody a particular vector, $\vec{V}(d_i), i = 1, n$.

The selection of terms for populating these vectors has a direct impact on the generated results, elaborated later in this section. Additionally, the frequency of a term in the vector can shadow the importance of any relevant term, with a lower frequency.

Accordingly, the importance of each word from the selected metadata is weighted by applying the TF-IDF algorithm [MaRS08, Ramo03].

6.2.1 Determining the key terms of a publication

Topic modeling, and extracting the key terms from a given text, is a very common and applied issue in IR, NLP, data mining, etc. Consequently, there is a large number of techniques and mixtures among them for performing in different environments. Such that the TF-IDF, Latent Semantic Analysis-LSA including the probabilistic attitude, Latent Dirichlet allocation-LDA are among the most popular for this purpose.

6.2 Vector Space Model approach

However, there are present several cases when the existing methods are modified or combined with other techniques, for the sake of computational cost or performance quality. Such an example is the mixture of Dirichlet Topic Models with Word Embedding for creating the *Ida2Vec* [Mood16]. On several other occasions, the application of external vocabularies and thesauri may be used for labeling documents with a set of controlled terms.

In our approach, we base the initial experiment on the basic TF-IDF approach, for having a better view and adjustments over the metadata terms and analyzing their role in the crosslinking process. However, in that case, the IDF value is not generated based on the corpus, but it was adopted from a general frequency of terms based on Google Books Ngrams (GNB). GNB presents a dataset of n-grams consisted of unigrams to 5-grams corpus [BrFr06, Norv13, Stef10]. In this work, we are focused on unigrams, i.e. individual words and their frequency in the corpus. Thus, locally we have saved a dataset consisted of 319 999 words of English language and their frequency of usage. Table 6.2 gives a short overview of some words and their frequency over that dataset. As expected, the word “*the*” is the most used with a 0.0393 frequency.

Table 6.2 The list of unigrams, the word and their frequency based on Google Books Ngrams

Word (w)	Frequency (f_w)
the	0.03933837507090550000
of	0.02236252533830050000
and	0.02210015761953700000
to	0.02063676420967820000
high	0.00058731326672819600
money	0.00032340969045539800
food	0.00030630268711382300
bank	0.00015568028973689900
taxes	0.00005725775413448000
inflation	0.00001456795810462590
leverage	0.00000687497277142300

6. Linking publications across different LOD repositories

In advance, before populating the vector $\vec{V}(p)$ with terms from the set of publications' metadata, several pre-processing steps are performed; such as removing punctuations, lowercase and encoding the data to Unicode character encoding (UTF-8). Additionally, the list of "stopwords" is applied for avoiding the iteration at table 6.2 for high-frequency words. After that, each word that becomes part of the vector is weighted by considering a very naive method. In the case when the word (w) is listed in the frequency dataset, its weight is determined by multiplying $(1 - f_w)$ with the term frequency. Otherwise, if the word is not part of that list, the weight remains to be calculated based on the metadata distribution. The equation below gives more details.

$$w_{weight} = \begin{cases} \log 10 \left(1 + \frac{tf}{n} \right) * (1 - f_w) & w \in W \\ \log 10 \left(1 + \frac{tf}{n} \right) & w \notin W \end{cases}$$

In the majority of experiments, a global unigrams frequency of words is applied, instead of generating corpus-based frequency, which is a common practice in TF-IDF implementation. An instance of these frequencies is shown in table 6.2. The only reason for this approach relies on avoiding the domain influence over the generated frequencies since we are aiming to crosslink interdomain information.

Let us consider the title of the publication "*Do inflation and high taxes increase bank leverage?*". After the preprocessing steps, the vector $\vec{V}(p)$ will contain the words *inflation*, *high*, *taxes*, *increase*, *bank*, and *leverage*. In this case, the overall number of words in the vector is denoted as n , $n=6$, while the frequency of the words in the document (i.e. title) is denoted as tf .

As shown in figure 6.6, for a given paper from the initial repository, the developed prototype makes it possible to adjust the relevance of each metadata component: the value can be increased or decreased by weighting the title(p_t), abstract(p_{abs}), keywords(K_p) and descriptors(D_p). The example shows that if we only consider the title of the selected publication (see figure 6.6-a), the words "leverage" and "inflation" are more crucial, whereas "high" is less important. This is because in general "high" occurs very often (based on table 6.2).

bank taxes
high leverage
inflation

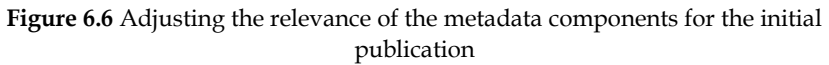


Table 6.3 The top-ten most important terms of a publication metadata based on a specific adjustment among them.

75

6. Linking publications across different LOD repositories

The top-ten most important terms regarding this metadata adjustment for this publication, are listed in table 6.3. Besides the fact that the term “high” appears seven times in these metadata, it is ranked as the sixth most important word. A better visual interpretation of these weights is shown in figure 6.6-b.

6.2.2 Measuring the similarity among publications

The similarity among publications, i.e., vectors of concepts, is measured as the deviation of angles between each document vector, by using the Cosine Similarity. Thus, iteratively we measure the similarity between metadata of our initial publication with the metadata of publications from the target repository, $\text{sim}(p, d_i)$, for $i = 1, n$. As shown in figure 6.6, the combination of the metadata is crucial for determining the weight of the terms in the initial publication. The proper selection can be seen as the right bait for successful “fishing”. Different combinations among these parameters would result in a different list of retrieved publications from the targeted repository. The impact can also be seen in the generated results.

In another study [HaLT14], considering different cases, different combinations of these metadata also led to good results. For this purpose, we conducted heuristic evaluations when analyzing the impact of each element. In the absence of any golden rule, as the most determinant combination we have perceived the combination of all the available metadata elements by doubling the title, $(2p_t, p_{abs}, K_p, D_p)$. The title is far representative, since the author tends to include the key terms regarding the subject.

6.2.3 Experimental setup of VSM approach

We have evaluated 57 EconStor publications, through the developed prototype. Thus, after triggering a title from EconStor, the system retrieves an ordered list of most similar publications from other repositories, in this case, AGRIS. As can be seen in figure 6.7, the prototype generated values for several parameters. Thus, $\#c$ represents the number of common

6.2 Vector Space Model approach

descriptors in both sites. $\#w$ is the number of common words among these publications. Tcs represent the cosine similarity measured only on titles, while $simCS$ the cosine similarity measured with all the defined components, i.e. $sim[(2p_t, p_{abs}, K_p, D_p), (2d_t, d_{abs}, K_d, D_d)]$.

From the generated results in figure 6.7, the prototype shows that the first retrieved publication has 0.3370 cosine similarity with our publication, while among titles the similarity is zero, since there are no common words on both sides. The value of one in the parameter $\#c$ represents the intersection of common words between our publication and this one. The prototype indicates that the average number of tokens from the initial publication is about 72, while at the targeted part this number goes to 79.

However, the frequency of tokens inside the metadata is crucial for scoring results. The word “inflation” in fact appears eight times in our selected publication at the initial repository and 20 times in the first ranked paper from the target repository. Conversely, the number four has only two words in common i.e. the word “inflation” and “bank”, and a small number of other noisy words. The number of the equivalent descriptors used for describing a paper in both repositories generally is one; except for a few cases, the publication is described by two or more descriptors in both repositories.

For determining the relevance of the retrieved publications, human evaluations are done on the top-ten ranked results. These evaluations are done by analyzing and comparing the titles [Resn61] and continuing with the abstract using the possibility for full-text reading.

Actually, evaluators have been asked the question: *How would you evaluate the relevance of the top-ten retrieved publications regarding the selected publication?*

To each of the top-ten retrieved publications is assigned a value of **i** for **irrelevant**, **s** for **somehow relevant** or **r** as **relevant**. Therefore, considering the example in figure 6.7, publication number three is evaluated as irrelevant and seven others as somehow relevant, while only two of them are depicted as closely related to the initial publication. However, in another example, “Food prices and political instability” inside the top-ten, five are identified as irrelevant, four somehow and only one as relevant.

6. Linking publications across different LOD repositories

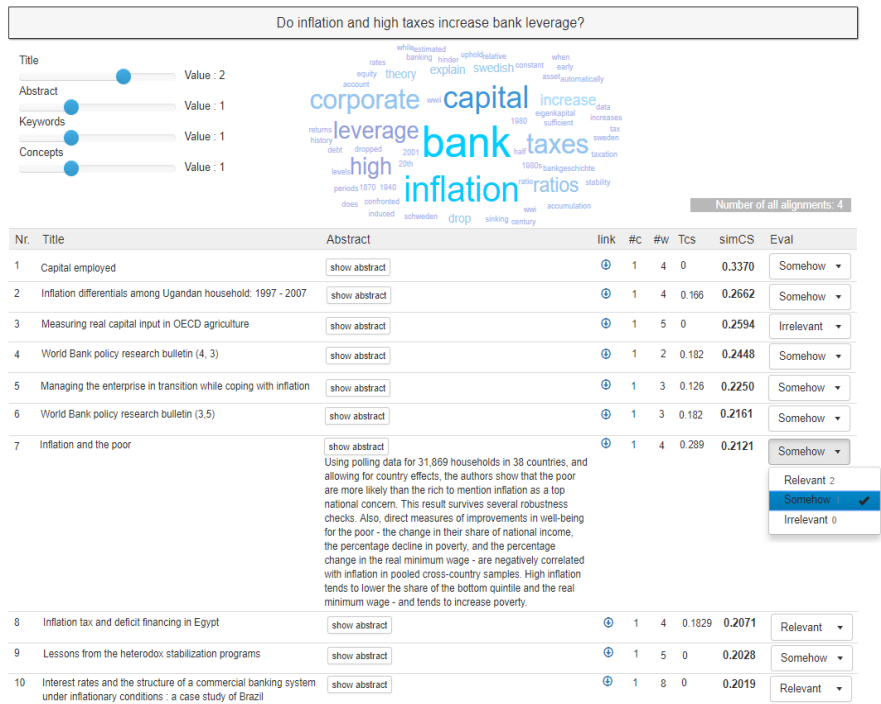


Figure 6.7 The combination of metadata components from a scientific paper for retrieving recommended publications from other repositories

The precision (i.e. the list of relevant documents) is improved by factorizing the title in the scoring and ordering. A simple experimentation, by performing the ranking as the average of *title* and *all other metadata* i.e., $avg(Tcs, SimCS)$, shows significant improvement regarding the number of relevant publications in the top-ten. However, this has negative implications for the relevant publications with no meaningful title. In that case, several relevant publications will not be highly ranked. The tenths ranked publication from figure 6.7 that is evaluated as relevant will not be in top-ten since the zero value in *Tcs*. As a result, the role of the abstract and other components such as keywords or descriptions is crucial when the title is not subject representative (of the type “*What next?*” or “*Lessons learned*”).

Hence, when the title does not contain common terms with the publication's metadata, e.g. "*Capital employed*" in figure 6.7, VSM fails to calculate any similarity.

The count-based approach with TF-IDF and Cosine Similarity generates satisfactory results for retrieving relevant publications from other repositories when a satisfactory amount of metadata is provided. Especially when the intersection between the compared documents results in common words [HaLT14]. Despite that, we have identified several weaknesses showed with this approach, in several directions. The next section highlights some of these problems.

6.2.4 Limitations of VSM

The main issue with this approach is that it is strictly related to the intersection of common words among compared documents. Such that, a simple morphological variation between words deviates the result. The attempt for achieving uniform words, i.e. converting to singular, or by applying stemming or lemmatization, show improvements. However, we need to be very careful with this process, since the evaluations show that in several cases the stemming or lemmatization can be so "aggressive" by changing a word roughly. In figure 6.7, our title with the title of publication number ten initially generates zero similarity (Tcs). After the stemming process, the word "*bank*" will be matched but not "*inflationary*" since it is stemmed like "*inflationari*".

The semantic interconnection between words or the context of use is not taken into account, as we cannot find any similarity or relatedness between words such as "*bank*" and "*credit*". This implies that a large number of relevant publications might not be on the top. The application of external vocabularies such as WordNet, for the availability of synonyms about the given words, even more, complicates the process. The variety of synonyms for a single word broadens the result by making it too far from the initial publication.

This approach repeatedly shows those irrelevant terms to be highly ranked. Let's take the publication titled "*Food prices and political instability*", based on the combination $(2p_t, p_{abs}, K_p, D_p)$, the word "*food*" becomes

6. Linking publications across different LOD repositories

dominant. This results in compromising outcomes, i.e., recommending semantically distant publications to that publication. In this case, as a first ranked publication, we retrieve *“Food Security in Older Australians from Different Cultural Backgrounds”*. Therefore, the right combination of metadata terms for this purpose is very experimental.

Another point worth mentioning is that this approach shows unsatisfactory results when measuring the similarity of vectors with only a few terms in them, such as the similarity between titles of publications. As one of many examples that show the weakness of these approaches when relying on short texts, is the similarity between these two titles *“Do inflation and high taxes increase bank leverage?”* and *“Lessons from heterodox stabilization programs”*, which results in zero.

6.3 Word Embedding Approach

Section 6.2 centered on the use of TF-IDF and CS for measuring the similarity among publications and realizing a ranking according to the similarity score. Based on this, in general, TF-IDF and CS do not offer much for achieving a completely automated process for measuring the semantic relativeness among the initial and retrieved publications [BaDK14]. Having this, we have explored several other approaches for finding an optimal solution that includes the semantic component for similarity measurement and ranking. The Latent Semantic Analysis (LSA) [DDFL90] or Latent Dirichlet Allocation (LDA) [BINJ03] are an option in this direction. However, based on the evaluations in several studies, these approaches do not offer the best solution for our cases [BaDK14, KiWL16, PeSM14]. According to this, we are focused on neural word embedding as one of the most promising approaches in the NLP.

6.3.1 Training and Building the Model

The experiments in this section are based on the Gensim package, which is a Python implementation of the Word2Vec model [ŘeSo10]. Gensim provides significant optimization regarding the computational speed,

6.3 Word Embedding Approach

which overpasses even the native C implementation. Currently, there are several pre-trained models on different datasets, such as Google News, DBpedia, and Freebase. Considering the specificity of the domain, we prefer to train our own word vectors for deploying the experiments.

Our model is trained on a text corpus for generating a set of vectors, which are word representations of words in that corpus. Through a SPARQL query, we retrieve all the titles, abstracts, and keywords of 37 917 publications from EconStor. Since Gensim's Word2Vec expects a sequence of sentences as input, several preprocessing steps are performed at the corpus, such as conversion to utf8 Unicode, lowercasing, removing numbers and punctuations. Finally, the model is trained on the corpus of 12 329 307 raw words and 683 937 sentences. Before the training process, several hyper-parameters are determined in concert to the training speed and quality. Based on our dataset size, every word in the corpus is considered with a window value of five. The dimensionality space of the words inside a vector is set to 300, which means that each word is represented with 300 most similar words in that vector. More words in a vector mean better quality, although a bigger dataset must be used. The hierarchical skip-gram architecture is used for training the model in a laptop with i5 CPU 1.7 GHz, 8 GB RAM memory. Surprisingly, the time it took was 129.7 sec, far beyond our expectations. This is the main model where we have based the experiment, otherwise, we have trained and tested many others by changing the hyper-parameters or even the corpus.

6.3.2 Analyzing the Model

This section presents the investigation of the learned model. We performed several analyses on top of the trained model in section 6.3.1. One of the most interesting analyses regarding the word representation approach is about finding the set of related words based on a particular entered word. For instance, regarding the economic domain of the trained corpus, we are interested to see what the model learned about the concept "*inflation*", as a purely economic concept, and the concept "*food*", as a general concept. Table 6.4 lists ten nearest terms that Word2Vec has calculated for these words.

6. Linking publications across different LOD repositories

Table 6.4 Top-ten most similar words based on the words “inflation” and “food”, generated through Word2Vec from our text corpus.

a. for the word “inflation”		b. for the word “food”	
Word	Similarity	Word	Similarity
output	.644	energy	.789
nominal	.611	agricultural	.786
volatility	.604	water	.767
gdp	.590	land	.756
aggregate	.570	crop	.701
persistence	.561	fuel	.694
macroeconomic	.543	transport	.694
price	.535	agriculture	.691
inflationary	.532	electricity	.690
forecast	.531	milk	.684

The generated results are very impressive. For example, the word “output”, “nominal” and “volatility” are ranked as the most similar to “inflation” with a degree of similarity .644, .611 and .604 out of 1. In fact, that value is more accurately to be denoted as the degree of relatedness among these concepts, rather than the similarity [FTRD16]. In general, all the listed words are intuitively very close to it. Moreover, a word is represented in relatedness to three hundred words, as defined by the training parameters. To our knowledge, it is almost impossible to generate such a result through dictionaries or thesauri. Hence, if we are referring to the STW thesauri described in section 3.3.1, the concept “inflation” is not represented with many meaningful terms, regarding the SKOS vocabulary. Even the usage of other external resources, such as WordNet synonyms, does not offer such an impressive set of related terms.

The trained model can be used for several other features of semantic language processing. Accordingly, there is a possibility to retrieve a list of most similar words by subtracting words from a given set of words. Thus, from a set of metadata, we have the possibility to include or exclude several concepts. For example, from the set of metadata concepts defined for a publication, we want to consider the terms “bank”, “oil” and “price” by excluding the term “food”. Therefore, based on this formula [(bank + oil +

6.3 Word Embedding Approach

price) - (food)], the trained model offers the term “currency” with .764 similarity, “liquidity” with .734 and “spreads” with .695. Such an implementation can be useful and determinant in the steps to populate the vector $\vec{V}(p)$ with terms from the publication’s metadata. Initially, the evaluations rely on a completely automatic metadata selection process for populating the vector. Moreover, in section 6.4 the user interaction with the metadata properties, in regard to different adjustments such as selection and weighting, is presented.

6.3.3 Experimental setup of Word Embedding approach

Based on the developed prototype, we have evaluated exactly the same 57’s EconStor publications, used in section 6.2.3. For each selected publication, the prototype retrieves and orders the most semantically similar publications from AGRIS. The process is the same as in section 6.2.3, however as can be noted from figure 6.8, in this approach we have introduced two more measurement components; $Tw2v$ which denote the Word2Vec similarity among titles, and $simW2V$ that is the Word2Vec similarity measurement among all the publication’s metadata, i.e. $sim[(2p_t, p_{abs}, K_p, D_p), (2d_t, d_{abs}, D_d)]$. The ordering is performed according to $simW2V$ scoring.

As expected, the implementation of the word embedding approach shows a different list of retrieved publications, compared to Cosine Similarity in figure 6.7. The results from the figure 6.8 make it obvious that the values generated through Word2Vec overcome those generated by CS. Figure 6.8 represents one of the depicted results from the evaluated publication, which is the same as in figure 6.7, “Do inflation and high taxes increase bank leverage?”. The results are shown in both approaches with two different sets of metadata

Firstly, the similarity degree between publication p and d_i is calculated only on titles, such as $sim(p_t, d_{t,i})$. As such, for the first retrieved publication on that list Word2Vec has generated a similarity of .5680, shown in $Tw2v$ column. The count-based implementation of Cosine Similarity gives 0 score between the same titles, shown in Tcs . This is one of many examples that

6. Linking publications across different LOD repositories

prove the ability of the word embedding approach to work even with a small amount of metadata.

In the same example, analyses are extended by including other metadata terms in the similarity calculations. Hence, from the EconStor publications the **title**(p_t), **abstract**(p_{abs}), **keywords**(K_p) and **descriptors**(D_p) are considered, while from the AGRIS publications the **title**(d_t), **abstract**(d_{abs}) and **descriptors**(D_d). The last two columns of figure 6.8 show the similarity among these metadata comparatively in both approaches, *simCS* and *simW2V*. By considering the first publication from figure 6.8, TF-IDF with CS generates .2019 similarity degree among them, while Word2Vec gives .8733. It is more than obvious the differences of generated results in both approaches. What is most important lies in the fact that WE reaches to rank on top publications that the previous approach could not. Therefore, the third-ranked publication through Word2Vec, manually judged as relevant (see figure 6.8), does not appear in top-ten retrieved publications in the first approach where only CS is applied (see figure 6.7).

Nr.	Title	Abstract	link	#c	#w	Tcs	Tw2v	simCS	simW2V	Eval
1	Interest rates and the structure of a commercial banking system under inflationary conditions : a case study of Brazil	show abstract	🔗	1	8	0	0.568	0.2019	0.8733	Relevant ▾
2	Inflation tax and deficit financing in Egypt	show abstract	🔗	1	4	0.1829	0.7481	0.2071	0.8644	Relevant ▾
3	Inflation and the rule-of-thumb method of adjusting the discount rate for income taxes	show abstract	🔗	3	8	0.1873	0.7125	0.1978	0.8565	Relevant ▾
4	Capital rising in the Baltic States: lessons learned and future prospects	show abstract	🔗	1	8	0	0.5081	0.1561	0.8523	Somehow ▾
5	Effects of tax incentives on long-run capital formation and total factor productivity growth in the Canadian sawmilling industry	show abstract	🔗	1	6	0	0.6921	0.165	0.8478	Somehow ▾
6	Future capital requirements should be studied	show abstract	🔗	1	6	0	0.6133	0.1589	0.8437	Irrelevant ▾
7	Returns, interest rates, and inflation: how they explain changes in farmland values	show abstract	🔗	1	5	0.1179	0.7174	0.1679	0.8335	Somehow ▾
8	Macroeconomic factors influencing lending rates	show abstract	🔗	1	6	0	0.5844	0.1268	0.8326	Somehow ▾
9	Dollarization and exchange rate fluctuations	show abstract	🔗	1	6	0	0.6729	0.1506	0.8299	Somehow ▾
10	Liberalising foreign investments by pension funds: positive and normative aspects	show abstract	🔗	1	9	0	0.5768	0.0571	0.8226	Irrelevant ▾

Figure 6.8 The similarity measurement is scored with cosine similarity and Word2Vec. The results are ordered based on Word2Vec similarity score. The relevance of the retrieved publications is evaluated manually.

6.3 Word Embedding Approach

The fact that word embedding overcome cosine similarity, regarding the score value, can not be adopted with automatism as the ultimate approach. Since the scores are used for ranking purposes, we have extended the human evaluation in both approaches, comparatively. Thus, the same as in the first approach, the top-ten retrieved publications are manually analyzed in order to determine the semantic relevance with the initial publication.

6.3.4 Limitations of Word Embedding approach

In the case when the word embedding model is trained on the corpus of one dataset, then the vocabulary of that corpus is embedded in word arrays. Such that, the usage of the model for measuring semantic similarity between two texts from different datasets is facing in a large set of “unknown” words. In our case the model is trained from the EconStor data, thus the Word2Vec has detected several missing words from the AGRIS when similarity measurement is calculated. We have ignored all the words that are not part of the trained model; however, this has implications in the generated results, i.e. the result to be generated on a few terms that cannot be representative for the publication from the non-trained corpus.

Using a model trained on a non-specific domain, such as Google News, decreases the number of missing words, given the wider range of covered vocabulary. However, the application of this model does not make evident any improvements regarding the relevance of the top retrieved publications. Building a model on top of the experimented datasets, the initial and the targeted repository is resulting in different distributions of semantically related words in arrays. Therefore, considering the combination of EconStor and AGRIS for building the model, Word2Vec gives more general context to a particular word, instead of closely related economic correlations. Thus, in this situation the most semantically similar words to “*food*” are listed, *seafood* 0.71, *foodstuff* 0.69, *grocery* 0.66, *restaurant* 0.651, *consumer* 0.642, *menu* 0.620, etc. As shown, there is a huge difference compared to the same word in table 6.4-b. By applying this kind of model, we are facing decreased performance in the task to determine the semantic similarity between two publications, according to human judgments. The embedding trained on specific domain corpora generates better results versus a more general model such as Wikipedia or Google News, for

6. Linking publications across different LOD repositories

specific related tasks [YSMB16]. In different scenarios, the combination of local and global context corpora in the learning process is productive for a more general word representation [HSMN12].

Word embedding is an unsupervised process, such that the selected dataset for training the model is crucial for the quality of the model. Therefore, the absence of terms in the training phase, word frequency and neighborhoods can be determining factors. Even the predefined hyper-parameters like the dimensions of the distributed words on arrays, the window size, negative samples or the minimum count, can play a role over the final model. Based on the performed experiment, we conclude that the word embedding knows to be very sensitive to these tuning parameters. Similar conclusions, regarding the tuned parameters, are noted in other works [FTRD16, HSMN12, LeGD15, SLMJ15, YSMB16, ZaCr16].

Recent trends are putting the focus on the combination of word embedding with other old-fashioned approaches (ex. LSA, BM25, TF-IDF), or with other different word representation methods (ex. Mikolov, Glove) [NGBM15] [KiWL16]. As claimed, such combination is resulting in better performance regarding the measurement of semantic word relatedness or even semantic text similarity. In [KiWL16] is proposed a combination of word embedding with BM25, or Word Mover's Distance (WMD) [KSKW15] for measuring similarity between a query and a document. While [BCBD16] shows an attempt to combine word embedding with TF-IDF information. The use of weighted centroids of word embedding with WMD to re-rank the retrieved documents is evident in [BrMA16].

6.4 UI integration and scholar involvement

The approach presented in the following section was originally published in our IEEE research paper [HaTo18], which includes the outcomes of a visual search interface application through users' involvement regarding the selection and adjustments of search terms.

Let's consider a scenario to find songs similar to what we just heard, with the opportunity to define some different features, such as lower rhythm and more dominant piano. The same can be said for movies, to find

similar movies to the ones we like, with fewer scenes of violence, more dramatic and a lot of mystery. In both scenarios, we seek a product that is recomposed of similar products by selecting features that we like or dislike. In the context of scholarly communication, on many occasions, the necessity for something similar is noticeable. Hence, let assume that we have found an interesting publication in our favorite DL, titled "*Globalization, brain drain and development*". Within the DL, we can get a list of recommended publications based on it, however, what if we prefer a list of publications, which are more related to "brain drain" rather than "globalization"?

The shown used cases have in the center the user behavior, i.e. the activity of the scholar in discovering publications. Namely, when a scholar refers to a DL, she initiates the search based on a set of terms, whose selection is very crucial for the results. Moreover, when an interesting publication is retrieved, the scholar's interest in other similar publications is obvious. Principally, almost every DL provides a list of feeds, i.e. recommending based on a selected publication. For example, Google Scholar¹⁶ offers the option "Related Articles", Mendeley¹⁷ has "Suggestions Based on This Article", EconBiz "Similar Items by Subject" while Elsevier's ScienceDirect¹⁸ offers "Recommended articles", etc. An in-depth overview for facilitating faceted search is provided by the EEXCESS¹⁹ project. However, from what we have observed, most of the existing approaches lack the opportunity for detailed adjustment of the searching parameters, with the purpose to customize the results. In addition to common layouts for specifying and narrowing down the results, when multiple functionalities are applied, the overload of the designs is obvious. This is especially evident when the scholar's search terms are extended through an external thesaurus or machine learning approach. Therefore, the scholar remains unaware about the presence of such terms in the query formulation and moreover why a particular publication appears in the result list. Within this context, our approach tends to introduce a balanced interface between simplicity and functionality, i.e. getting more with less effort.

¹⁶ <https://scholar.google.com>

¹⁷ <https://www.mendeley.com>

¹⁸ <https://www.sciencedirect.com>

¹⁹ <http://eexcess.eu/visualisations>

6. Linking publications across different LOD repositories

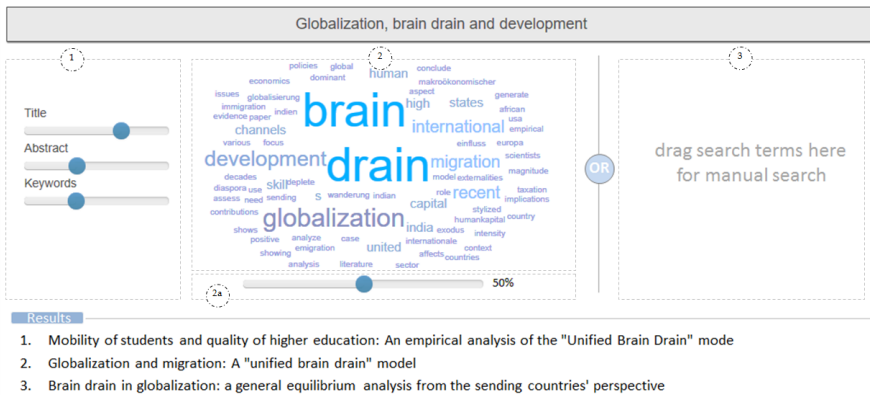


Figure 6.9 The proposed visual search interface

At this point, the main goal is to enable the scholar to redefine the list of recommendations based on the terms from a particular publication. To make this possible, the representative terms from the publication's metadata, such as title, abstract and descriptors/keywords, are extracted and visualized in a word tag cloud. Such that, the scholar get an instant and better overview of the topics and main concepts in that publication.

Hence, when the scholar selects a paper in a DL, the most important concepts based on the retrieved metadata from that publication are determined. Furthermore, the scholar can make adjustments to metadata components to define the selection of the concepts and also determine the weight of each concept in particular, in order to narrow down the results. The following sections present two main methods in that regard.

Similarly, as in the previous experiments, the approach is assessed with the content of the EconStor and AGRIS repository.

6.4.1 Automated Search

Figure 6.9 shows a scenario when through the provided interface the scholar selects a certain publication, and as an outcome, its metadata are projected in a word tag cloud. As depicted, the scholar's search interaction

6.4 UI integration and scholar involvement

is concentrated on three main areas, in order to advance her visual search and retrieve semantically similar publications. Through area 1, as denoted in figure 6.9, the scholar determines the set of metadata components to consider, by selecting among the title, abstract, and keywords. Thus, the scholar can include or exclude any of them, or specify the relevance by increasing their value over the sliders' interaction. As it is exemplified in figure 6.9, the title is factorized more in comparison to other components, therefore its terms will get more importance.

The outcome of the metadata combinations from the first area is instantaneously visible within the word tag cloud in area 2, while the application of TF-IDF, as denoted in section 6.2.1, achieve to highlights the most representative terms about that publication. From the same area, i.e., from 2a, the scholar can do the primary interactivity to generate a list of recommendations. Therefore, each interaction with the corresponding slider determines the number of concepts involved in calculating the semantic similarity for generating recommendations. Accordingly, given that the inclusion of all metadata elements (title, abstract, and keywords) produces a large set of terms, their selection in the similarity measurements leads to a more accurate list of recommendations. For this reason, the scholar can determine the number of terms to be considered, starting with the most emphasized ones. For that purpose, at this point, we are proving an intuitive and automated approach for selecting the terms.

Additionally to the automated selection of terms, the scholar can perform a manual selection, by choosing the finest combination to narrow the search. Therefore, by dropping the terms in area 3, the scholar can generate a specific query formulation with the set of terms involved in the search. We provide a detailed description of the customized search in the following section.

6.4.2 Customized Search

As described previously, the manual section of terms from the word tag cloud allows the scholar to perform a very refined formulation of the search query. Moreover, in addition to the assembly of the terms in area 3 (see figure 6.9), several other customizations can be implemented. At first, each

6. Linking publications across different LOD repositories

of the terms can get a distinct relevance by decreasing or increasing its weight on similarity calculations. A better overview of this feature is visible in figure 6.10. As represented, almost every selected term i.e., *globalization*, *brain drain*, *development*, is adjusted by the scholar, regarding their retrieval relevance. Such that, “*brain drain*” has been determined as the most crucial, while “*scientists*” as less important in that collection.

Furthermore, the set of terms can be extended by manually inserting new ones, besides the drag-and-drop option from the word cloud. This can be achieved through the “+” button, which generates a text box at the end of already existing terms. For example, by entering the word “*ict*” in figure 6.10, it directly becomes part of the searching set, with a default relevance.

The same figure also highlights some other important details regarding the extension of terms. As denoted there, each selected concept is accompanied by the symbols “t” in yellow color and “m” in red color. Through these two options, the scholar can enrich the provided terms with several others through the deployment of external thesauri (t) or with terms generated through machine learning techniques (m). Avoiding their presence in the recommendation retrieval can be achieved by pressing above the corresponding label. Such cases are evident in the terms “*development*” and “*scientists*”, where the faded silver color indicates the neutralization of the corresponding extensions.

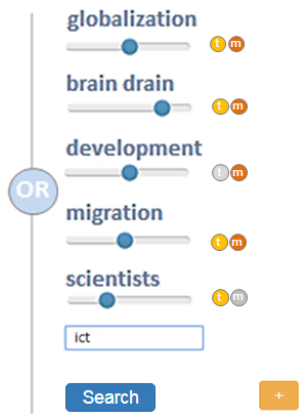


Figure 6.10 Customized search

6.4 UI integration and scholar involvement

Thesaurus Terms represent the terms suggested by the usage of an external thesaurus. Such an example can be the deployment of a general lexical database or even a domain-specific vocabulary. Hence, the WordNet implementation may help us get the synonyms or semantic relations for a given word [Mill95]. However, in our evaluations, we have adopted the STW thesaurus (see 3.3.1), taking into account the economic domain of our initial repository. Through the SKOS modeling scheme, STW enables hierarchical navigation between concepts, in narrowed, broadened, and related terms. In general, the approach is not limited to a specific thesaurus or vocabulary, but if it is close to the respective domain, it positively affects the accuracy of the retrieved results.

The example in figure 6.11 illustrates a list of suggestions related to the concept “*globalization*”, according to the STW thesaurus. For practical reasons, the number of suggestions is limited to ten. When the “t” option is enabled, the entire list becomes part of the information retrieval process. However, the scholar may exclude any of the terms, as has been done with “*transnationalization*” in the same figure, if she assesses it as unnecessary, or perhaps an outlier who may deviate the outcome. As explained before, there is a possibility to deactivate the entire list, as shown with the term “*development*” in figure 6.10.

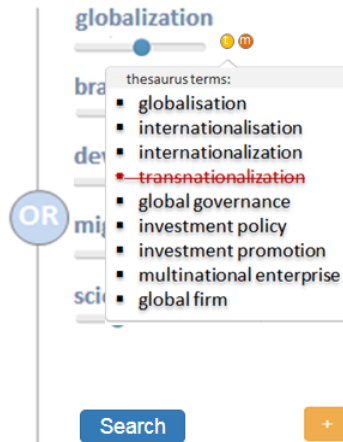


Figure 6.11 Additional thesaurus suggested terms

6. Linking publications across different LOD repositories

Machine Learning Terms are generated through the application of the machine learning methods, such as the word embedding approach. For that purpose, as described in section 6.3.1, a model has been trained and built based on the EconStor repository subset through the Word2Vec technique. Figure 6.12 gives the top-five most related terms considering the term “globalization”, relying on the already built model. It is worth mentioning that the usage of different models leads to completely different outcomes regarding the relatedness between terms, i.e., suggestions. For instance, the Google News model categorizes the following terms as most similar to “globalization”: *globalism*, *globalized*, *globalizing*, *globalization* and *capitalist_globalization* (not shown in figure 6.12).

All the features explained in the previous section, regarding the inclusion and exclusion of suggestions, apply here as well. The scholar can determine the presence of any concept or the entire list as a whole. Compared to thesaurus terms, the suggestions created by the machine learning approach are initially limited to five, with the possibility to expand the list to five more terms (by pressing the +5 option). Thus, in the case of "globalization" the following terms will be added to the list: *integration*, *liberalisation*, *deep*, *global* and *resilience*.

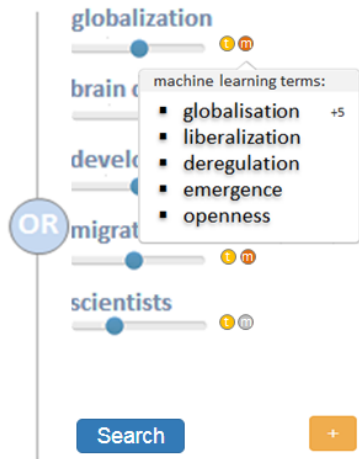


Figure 6.12 Terms suggested from the machine learning approach

6.4 UI integration and scholar involvement

The presented approaches, initially through the automated search, advancing to the customized form, represent significant facilitation of the search and retrieval process within or across DLs. The implemented visual search interface enables the scholar to operate with various functionalities and different sets of terms to a better query formulation and accuracy of the retrieved recommendations. However, the inclusion of a large number of terms, generated in different forms and sources, in several situations can lead to user uncertainty as to why a particular publication has appeared in the list of results.

An instance of such visualization is given in figure 6.13, especially when the customized search is applied. Hence, the black bolded text is related to the manually selected terms, from the word tag cloud to area 3; the concepts in red color originate from machine learning suggestion; and the text in yellow color characterizes the terms from the thesaurus suggestions. Such an appearance includes the title, abstract, and keywords. Moreover, through the mouseover event, the origin of the match is clearly indicated.

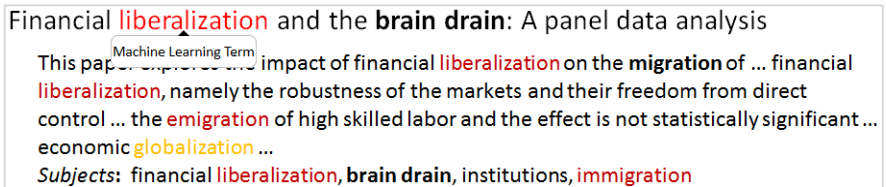


Figure 6.13 Retrieved results based on the visual search approach

Crosslinking-through author's disambiguation

*"Disorder created connections
-that is, resonance"*

Eric Abrahamson

Chapter 5 under section 5.3 expounds details about the proposed directions for achieving the crosslinking process regarding the data enrichment goal. Therefore, the second proposed direction is related to **author-centered metadata**. Hence, for a given author to find the correlations with other authors, publications or other related information by crosslinking data.

The idea:

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 author, of her or his co-authors, and additional bibliographic information of the initial author.

This chapter describes the approach of enriching the content of a DL with additional information from other repositories specifically regarding authors' related information. To that purpose, the main objective is to collect and crosslink author's bibliographical data, other publications, co-authors' relations, citation metrics, and everything else of interest for that author. Consequently, an extended profile will be created, as a combination of the repository data and the data found in other repositories.

For every author as part of a Digital Library, i.e. EconStor, 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. For achieving

7. Crosslinking-through author’s disambiguation

this goal, we extend our interest to other bibliographic repositories offered by several libraries and institutions. Of particular interest are the data which are presented in the form of Linked Open Data (LOD), as part of the LOD cloud [BHIB08, HaLT14, LaBT14, PaKS15]. As a test case, we target the following library and non-library sources: German National Library (DNB), Library of Congress (LoC), National Library of France (BNF), National Library of Sweden (KB / LIBRIS), DBpedia and WIKIDATA.

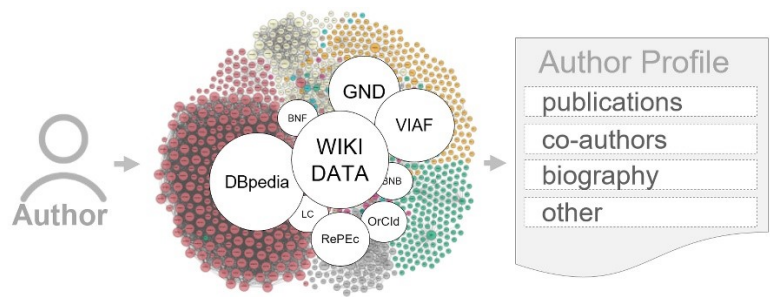


Figure 7.1 The author’s enrichment approach

However, retrieving the author’s details from other repositories remains to be a challenge. The author’s name ambiguity represents the major obstacle for direct information retrieval about a given author from corresponding repositories. In cases when an identifier is assigned to an author, such as ORCID, RePEc, GND, or something similar, the data crosslinking process, namely, retrieving the data from another repository for this author, is of reduced complexity. In addition, if the author is identified with a particular id, e.g. RePEc, the use of authority linking hubs, such as WIKIDATA or VIAF, may also reveal some other identifiers associated with this author. In the cases where the author does not have any type of identifier, there is always a doubt about whether we are pulling information for the right author, and it is a barrier to further data enrichment. For that purpose the deployment of author disambiguation is compulsory. Therefore, the creation of the author profile is preceded by the process of author name disambiguation, which is described in detail later in this chapter. This is presented as an inevitable need to achieve satisfactory and acceptable results. Therefore, the creation of a general profile for each author in a DL can serve for two purposes:

- to enrich the search result with several other information related to that author, and generating a wider profile about the author by integrating the most relevant harvested information about her.
- to resolve author name ambiguities through the unique identification of the same author written in different ways or the same name referring to different authors. This can contribute to clustering author names alternatives under one identification. Thus, we are increasing the precision of the retrieved publications, when a scholar is looking based on that author.

7.1 Author's name disambiguation

The process of correct author identification in different repositories is related to the challenge of the author's name ambiguity, when determining if two or more references correspond to the same person [EIV07, LGCF08, SGLF14]. For instance, an author can be represented with different spellings i.e., name alternatives, in several bibliographic repositories, or different authors can share the same name, which increases the complexity of the data crosslinking process.

As an example, we would like to find as much information as possible about an EconStor author by harvesting other repositories. However, in almost every case we encounter situations in which the same author appears with different name variations, such as **Adam Smith; Smith, Adam; A. Smith-; Smith. A.; Smith, Adam, 1723-1790; Смит, Адам, 1723; Smith, Adam T. ; and Smith, Adam, 1930**. Besides, there could be different authors, all named **Adam Smith**, or with related name labels. In principles, a similar problem concerns the metadata about titles of publications which can vary across different repositories.

The process of author disambiguation, as we have described in a previous publication [HaRT15] and redefined in the following chapter, in addition to the name, makes it necessary for the implication of several other metadata from the initial repository. Hence, from the set of author's centered metadata, explained in section 5.5, for a given author **a**, we are considering:

7. Crosslinking-through author's disambiguation

- the full name of the author \mathbf{a} ,
- the list of publications, denoted as P_a ,
- the list of co-authors of the author \mathbf{a} , denoted as A_a , and
- the list of publications of co-authors of \mathbf{a} denoted as \bar{P}_a .

In view of this set of metadata, we can target any of the proposed repositories, for harvesting information regarding the author \mathbf{a} . However, there are several initiatives that are already working for authority profiles, which offer the possibility of using them in this context.

Currently, there are present several efforts for generating authority profiles for uniquely identifying resources and researchers. As the most appropriate approaches that would facilitate the disambiguation of authors and which are used as a "bridge" for retrieving accurate information from other repositories, we emphasize GND, ORCID, VIAF, VIVO, RESERCHERID, and OpenID. Section 2.3 provides more details about these initiatives. In our work, we consider VIAF with the most usage relevance. Therefore, we utilize it as a "bridge" for the disambiguation process and crosslinking different bibliographical repositories.

Based on what we have encountered so far, repositories or DLs have different states regarding disambiguation quality. Hence, there are cases when a particular repository is entirely ambiguous, which means none of the authors is identified with any type of locally or globally identifier. In such a situation, there is almost impossible to distinguish and cluster the entire list of publications/co-authors related to that author. In such a situation, a record-based approach is followed, i.e. as input in the disambiguation process is considered only the title of that record and co-authors belonging to it, if there is any. Also, the presences of other persistent identifiers in that record, such as ISSN, DOI, SSRN, HANDLE, etc., are of huge benefit. After several authors' iterations in the disambiguation process, such that explained in section 7.2, the repository will achieve a partial disambiguation level. It means that several records will contain authors with a preferred globally identifier (e.g. VIAF ID, GND ID, ORCID, etc.). Moreover, in a partially disambiguated repository, the disambiguation process may continue and be divided into two main parts. Thus, a completely altered process can be performed locally, by analyzing only the metadata inside the repository, such as the list of publications where that author is already identified, the co-authors' graph, and the

7.2 Identifying Authors in VIAF

persistent identifiers on the corresponding records. In the cases when the local disambiguation steps are incapable to perform any results (due to the lack of information), the use of an external resource is considered over again. As noted before, as the most reliable resources for this purpose we have adopted VIAF.

The accurate identification of a particular author from a repository, i.e., EconStor, with the corresponding author (cluster) in VIAF, is just a straightforward step. As explained in session 2.3.1, there are also evident some weaknesses in the clustering authors to a specific heading. Some of these anomalies can be a wrong publication, co-author or a reference to any external resource. However, the most widespread problem is the large number of retrieved clusters searching with a particular author's name. Therefore, obtaining the accurate cluster will ensure a set of persistent identifiers, i.e., the VIAF ID and the IDs of corresponding sources contributing to that cluster., such as DNB, LoC, BNF and LIBRIS (fig.7.2). For this purpose, we are proposing an algorithmic approach by considering and comparing the set of metadata from the initial repository with the metadata found inside of each cluster.

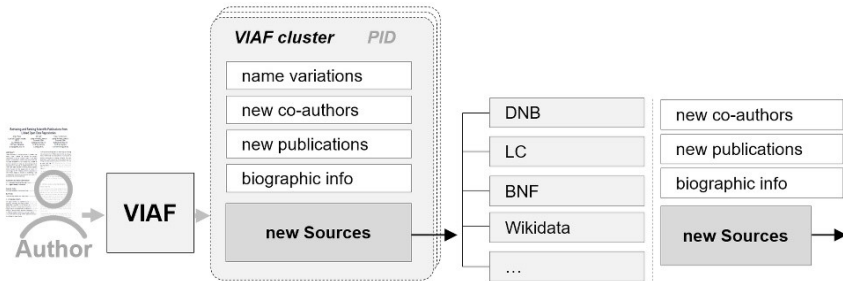


Figure 7.2 The overview for harvesting author's data

7.2 Identifying Authors in VIAF

As mentioned in the previous section and underlined in part 2.3.1, each search in VIAF can result in several records (clusters, headings) that match the name of an author. Therefore, one of the main challenges in this step is

7. Crosslinking-through author’s disambiguation

to assess the accuracy of each retrieved cluster, by analyzing and comparing author-related data from the initial repository with those found in each VIAF cluster. For this purpose, in addition to the metadata from the initial repository, we are considering several metadata components from each retrieved cluster c_j ($j=1$ to some hundred clusters), such as:

- Author’s name variations (alternatives), denoted as A_{c_j} ,
- Birth year and death year, C_{byear} , C_{dyear}
- List of publications in that cluster P_{c_j} ,
- List of co-authors in that cluster \hat{A}_{c_j} ,
- Publications from other sources assigned to that cluster \check{P}_{c_j} .

For determining the matching degree between the author **a** and the extracted clusters c_j , several data mining techniques are implemented. Therefore, by adopting different vector space algorithms, there is proposed an algorithmic approach. With the highest priority, we use the Cosine Similarity (CS) for measuring the similarity between publications, while we apply Levenshtein distance and Jaro distances for the similarity of author names. The algorithm we propose follows ideas from the process of name deduplication and address information [BiMo03].

We start by defining the metadata for the publications in our initial repository. These metadata are described in detail in section 5.2. In the very beginning, the process starts by using the VIAF API for identifying a particular author. Each retrieved cluster is analyzed iteratively according to four proposed steps, as emphasized in figure 7.3.

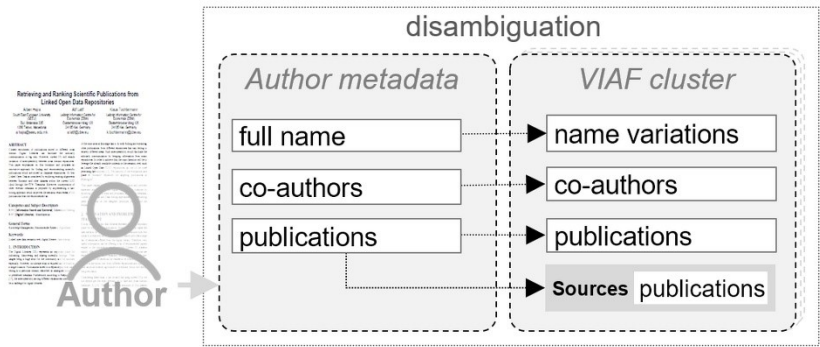


Figure 7.3 Identifying authors in VIAF

7.2 Identifying Authors in VIAF

Table 7.1 The calculated variables to assess the VIAF cluster accuracy match

Variable	Description
w_{ac}	the weight as results of similarity between author a with name versions in the cluster
w_{ac}^{\wedge}	the weight as results of similarity between co-authors of a with co-authors in the cluster c_j
w_{pc}	the weight as results of similarity between publications of a with publications in the cluster c_j
w_{pc}^{\vee}	the weight as results of similarity between publications of a with publications if libraries contributors in the cluster c_j

For each step in the process, there is calculated weight as a similarity degree between the corresponding metadata components from our author **a**, and the corresponding element from the VIAF cluster. The value of these weights is assigned to particular variables, as showed in table 7.1.

7.2.1 Author's name versus alternatives within a cluster

Each VIAF cluster consists of several name alternatives for the same author. Section 5.5.1 gives more details about this, by visualizing the cluster of the author "Smith, Adam". Therefore, the similarity measurement is calculated between the author's name from our initial repository with the alternatives within each cluster. In cases when at least one full match is found, a weight of 0.5 is assigned to the variable, denoted as w_{ac} . In detail, the similarity check is done only in the context of the author's 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 **a**.

$$sim(a, a_i^{cj}), i = 1, n; j = 1, k; \quad (7.1)$$

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.

7. Crosslinking-through author's disambiguation

7.2.2 Author's publications versus clusters publications

Another similarity measurement is done between publications that an author has in the initial repository with the publications found in the VIAF cluster. With P_a is assigned the set of all publications that this author has in our repository, while with 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, where each publication is presented as an array of words, 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, for a given publication from the initial repository, $p_e^a \in P_a$, $p_e^a = \{t_1^{pi}, t_2^{pi}, t_3^{pi}, \dots, t_k^{pi}\}$ and one from the VIAF cluster $p_f^{cj} \in P_{cj}$, $p_f^{cj} = \{t_1^{cj}, t_2^{cj}, t_3^{cj}, \dots, t_m^{cj}\}$ we have the similarity:

$$\text{sim}(p_e^a, p_f^{cj}), e = 1, k; f = 1, m; k, m \geq 3; \quad (7.2)$$

In this case for each comparison, a specific weight is assigned according to the calculated similarity value, denoted as w_{pc} . Based on the performed experiments, two main thresholds are defined.

Therefore, the variable of w_{pc} gets the weight score of 0.5 in cases when the similarity among the compared titles is between 0.6 and 0.9. However, this applies only to the instances when the number of tokens in titles is higher than three (considering equation 7.2). In this way, we want to avoid any unrealistic similarity score when in the comparison takes place a very short title. For every measurement that generated a similarity degree above 0.9, the value of w_{pc} is equal to 2. These values are set based on our preliminary analysis, which showed that lower thresholds and less than three terms in the title, resulted in inaccurate matching.

In a large number of cases, titles in the initial repository can be distinguished from them in a VIAF cluster with only one punctuation mark. Therefore, before performing the similarity algorithm, the cleaning and formatting of the data are conducted, such as: removing punctuation, eliminating "stopwords", lowercase and encoding the data to Unicode character encoding (UTF-8).

7.2.3 Co-authors versus co-authors in the cluster

The next step is related to the list of co-authors that our author has in the initial repository, comparing to 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 a has at least one common publication, while $\hat{A}_{c_j} = \{\hat{a}_1^{c_j}, \hat{a}_2^{c_j}, \hat{a}_3^{c_j}, \dots, \hat{a}_n^{c_j}\}$ is the set of co-authors in a particular VIAF cluster c_j . In this case, as it is explained in (7.3.2), each co-author from A_a is compared with each co-author from \hat{A}_{c_j} .

$$\text{sim}(a_e^a, \hat{a}_f^{c_j}), e = 1, k; f = 1, m; \quad (7.3)$$

At least one match, $A_a \cap \hat{A}_{c_j} \neq \emptyset$, can be significant proof that our repository and the cluster have a common co-author. In that case the variable $w\hat{a}_c$ will get a weight of 2 for each iteration based on CS. Having more than one match increases the evidence that it is the required cluster. Another similarity metric for names is applied based on the Jaro-Winkler metric. In this case, the similarity is calculated according to the characters and the $w\hat{a}_c$ weight is only 0.5. The threshold for names calculated by CS remains 1.0, while for Jaro-Winkler it will be above 0.9.

7.2.4 Author's publications versus publications from sources

The final check is related to the list of publications extracted directly from the sources (libraries) that this cluster has aggregated. The set of publications retrieved from the libraries that belong to the cluster c_j , is denoted with \check{P}_{c_j} . For example, if the German National Library (DBN) has its records in that cluster, there are measured the similarity between them and publications from our repository, $p^a \in P_a$ with $\check{p}^{c_j} \in \check{P}_{c_j}$. For each comparison, a weight of 0.5, i.e. 2, is assigned to the variable $w\check{p}_c$, absolutely in the same manner as in the step (7.3.2).

$$\text{sim}(p_e^a, \check{p}_f^{c_j}), e = 1, k; f = 1, m; k, m \geq 3; \quad (7.4)$$

7. Crosslinking-through author's disambiguation

7.3 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 (7.2.1), (7.2.2), (7.2.3) and (7.2.4) the weight is calculated iteratively with equations (7.1), (7.2), (7.3) and (7.4). Appendix A provides the complete algorithmic approach regarding the measurement.

The accuracy between the explored cluster and the initial author from our repository is determined based on the variables, w_{ac} , w_{pc} , $w\hat{a}_c$, $w\check{p}_c$. Therefore, in the case when the sum of any of these combinations $w_{pc}+w_{ac}$, $w\hat{a}_c+w_{ac}$, or $w\check{p}_c+w_{ac}$ is resulting in greater or equal to 2.5, the cluster is considered as "correct". While, between the values 1.5 and 2.5, the cluster is denoted as "maybe", below that value, the cluster is considered as "incorrect".

7.4 The experimental setup

For having a clear view of the applied approach, a prototype is developed. It offers an automatic approach to identify an author in VIAF and simultaneously extend her profile in the initial repository with other information. Thus, for a certain author, we are consuming the VIAF API, since VIAF strongly recommends its usage for up-to-date information. Below is shown the way of its usage, where the author's full name from the initial repository is considered as input, {name}.

VIAF API:

<http://viaf.org/viaf/search?query=local.personalNames+all+%22'{name}.'%22&sortKeys=holidayscount&maximumRecords=100&&httpAccept=application/rss%2bxml>

Since the output of the above API results in an XML structure, for each retrieved cluster that matches the corresponding author name, we are parsing the data inside it. The parsed content is similar to the data presented in figure 5.6 and table 5.4. Hence, the syntax for that purpose by including the namespaces is listed in the following.

7.4 The experimental setup

```
$xml->registerXPathNamespace('owl', 'http://www.w3.org/2002/07/owl#');  
$xml->registerXPathNamespace('rdf', 'http://www.w3.org/1999/02/22-rdf-syntax-ns#');  
$xml->registerXPathNamespace('ns2', 'http://viaf.org/viaf/terms#');
```

```
[Aij]          $nodes_AuthName = $xml->xpath('//ns2:mainHeadings/ns2:data/ns2:text');  
[Sij][Pij]    $nodes_source = $xml->xpath('//ns2:sources/ns2:source');  
[P̂ij]         $nodes_titles = $xml->xpath('//ns2:title'); //  
[Cbyyear]      $nodes_birth = $xml->xpath('//ns2:birthDate');  
[Cdyear]       $nodes_death = $xml->xpath('//ns2:deathDate');  
[Âij]         $nodes_coAuth = $xml->xpath('//ns2:data[@tag="950"]/ns2:text');
```

Accordingly, for each VIAF cluster the author name's alternatives, publications, co-authors, birth and death year, and including the identifiers of the other sources that composing this cluster are parsed separately.

Further, in addition to the cluster's content, we extend the range of data by considering and analyzing the aggregated sources within it, such as DNB, LoC, ISNI, and SUDOC. Each VIAF cluster offers these records in several formats such as MARC-21 record, VIAF Cluster in XML, RDF record and the links in JSON. We rely on RDF records, to perform a detailed exploration of each of the referred resources. To this end, particular interest has been given to data from the DNB.

7.4.1 The prototype examples

Through the developed interface, we are able to implement and analyze the approaches described above, with a visual overview of each step. Let's take a concrete example, by selecting a particular author, i.e. "**Sims, Christopher A**", in EconStor. As result, the prototype will provide various information about him that can be seen in figure 7.4, such as:

- the list of all publications,
- co-authors and
- co-author's publications.

All these data originate from the initial repository. Therefore, the co-authors' publications, which are sharing the co-authorship with the selected author, are noted with red color in bolded style.

7. Crosslinking-through author’s disambiguation

ECONSTOR | Authors

PROTOTYPE | Cross-linking author's information from other repositories.

select an author

Sims, Christopher A.

1. Bayesian methods for dynamic multivariate models (1996) [Link Econstor](#)

2. Does monetary policy generate recessions? (1998) [Link Econstor](#)

3. Error bands for impulse responses (1995) [Link Econstor](#)

4. Fiscal Aspects of Central Bank Independence (2001) [Link Econstor](#)

5. MCMC method for Markov mixture simultaneous-equation models: a note (2004) [Link Econstor](#)

6. Methods for inference in large multiple-equation Markov-switching models (2006) [Link Econstor](#)

7. Rational inattention: a research agenda (2005) [Link Econstor](#)

8. Were there regime switches in U.S. monetary policy? (2004) [Link Econstor](#)

9. When does a central ban's balance sheet require fiscal support? (2014) [Link Econstor](#)

1. Zha, Tao

A Gibbs simulator for restricted VAR models(2000) [Link Econstor](#)

Assessing simple policy rules: a view from a complete macro model(2000) [Link Econstor](#)

Bayesian methods for dynamic multivariate models(1996) [Link Econstor](#)

Conditional forecasts in dynamic multivariate models(1998) [Link Econstor](#)

Confronting model misspecification in macroeconomics(2012) [Link Econstor](#)

Do credit constraints amplify macroeconomic fluctuations?(2010) [Link Econstor](#)

Error bands for impulse responses(1995) [Link Econstor](#)

Indeterminacy in a forward-looking regime-switching model(2007) [Link Econstor](#)

Land prices and unemployment(2013) [Link Econstor](#)

Land-price dynamics and macroeconomic fluctuations(2011) [Link Econstor](#)

Learning, adaptive expectations, and technology shocks(2008) [Link Econstor](#)

Likelihood-preserving normalization in multiple equation models(2000) [Link Econstor](#)

Liquidity premia, price-rent dynamics, and business cycles(2014) [Link Econstor](#)

MCMC method for Markov mixture simultaneous-equation models: a note(2004) [Link Econstor](#)

Markov-switching structural vector autoregressions: theory and application(2005) [Link Econstor](#)

Methods for inference in large multiple-equation Markov-switching models(2006) [Link Econstor](#)

Minimal state variable solutions to Markov-switching rational expectations models(2010) [Link Econstor](#)

Modest policy interventions(1999) [Link Econstor](#)

Sources of macroeconomic fluctuations: A regime-switching DSGE approach(2010) [Link Econstor](#)

Structural vector autoregressions: Theory of identification and algorithms for inference(2008) [Link Econstor](#)

The conquest of South American Inflation(2006) [Link Econstor](#)

The dynamic striated Metropolis-Hastings sampler for high-dimensional models(2014) [Link Econstor](#)

2. Waggoner, Daniel F.

A Gibbs simulator for restricted VAR models(2000) [Link Econstor](#)

Asymmetric Expectation Effects of Regime Shifts and the Great Moderation(2007) [Link Econstor](#)

Confronting model misspecification in macroeconomics(2012) [Link Econstor](#)

Density-conditional forecasts in dynamic multivariate models(2010) [Link Econstor](#)

Generalizing the Taylor principle: Comment(2008) [Link Econstor](#)

Indeterminacy in a forward-looking regime-switching model(2007) [Link Econstor](#)

Inference based on SVARs identified with sign and zero restrictions: Theory and applications(2014) [Link Econstor](#)

Likelihood-preserving normalization in multiple equation models(2000) [Link Econstor](#)

Methods for inference in large multiple-equation Markov-switching models(2006) [Link Econstor](#)

Minimal state variable solutions to Markov-switching rational expectations models(2010) [Link Econstor](#)

Normalization in econometrics(2004) [Link Econstor](#)

3. Del Negro, Marco

Aggregate unemployment in Krusell and Smith's economy: a note(2005) [Link Econstor](#)

Asymmetric shocks among U.S. states(2000) [Link Econstor](#)

Country versus Region Effects in International Stock Returns(2003) [Link Econstor](#)

Inflation in the great recession and new keynesian models(2013) [Link Econstor](#)

Monetary policy analysis with potentially misspecified models(2005) [Link Econstor](#)

On the fit and forecasting performance of New Keynesian models(2004) [Link Econstor](#)

Policy predictions if the model doesn't fit(2004) [Link Econstor](#)

The FRBNY DSGE Model(2013) [Link Econstor](#)

The forward guidance puzzle(2012) [Link Econstor](#)

Time-varying structural vector autoregressions and monetary policy: A corrigendum(2013) [Link Econstor](#)

When does a central ban's balance sheet require fiscal support?(2014) [Link Econstor](#)

Search about this author

Figure 7.4 Initiating a search for a particular author.

106

7.4 The experimental setup

Considering the provided set of data, the prototype automatically checks, disseminates, and selects the best match of the VIAF authority clusters that match the name of the author. Consequently, for this author, the prototype has found five clusters in total, of which the first one is depicted as the correct match. Here is worth mentioning that VIAF regularly is updating the headings, therefore at different times, different results can be shown. Figure 7.5 gives exactly the view of the correct cluster.

In this cluster, in addition to the similarities between the author's name and the offered alternatives, similarities are also found between publications, co-authors, and other publications extracted from the corresponding sources that contribute to this cluster. Thus, figure 7.5 shows that the cluster contains 18 publications and nine co-authors related to the selected author. From the list of publications, the prototype has highlighted four publications with a similarity score of 100% to the publications from the initial repository, based on the calculations explained in section 7.2.2. Concurrently, into the same cluster, there are underlined two co-authors of "Sims, Christopher" (number 1 and 2) with 100% match, who are also co-authors in our repository.

Furthermore, there is a total of 14 libraries or institutions ("Sources" in Fig. 7.5) that contribute to this cluster. A possible assessment of these sources would enforce the matching degree, especially if we can identify publications that are not yet part of the cluster. For instance, two publications in German National Library are 100% similar to what we have in our repository (see "Other links" in Fig. 7.5). Nevertheless, in this case, the result is excluded from the overall calculation because the same publication appears in the cluster's publications, based on $p^a = p^{c,j}$ (publications 6 and 14 in Fig. 7.5). In general, all these elements provide evidence that this cluster is an accurate match for the selected author.

Figure 7.6 depicts one of the clusters, which the prototype has assessed as an inaccurate match for the author "Sims, Christopher". As can be seen, the calculated values are below all predefined thresholds for each required point. However, as noted in section 7.3, the prototype also assesses with "maybe" all clusters for which certain evidence is missing to be classified as correct or incorrect. Figure 7.7 represents an instance of such a case. Therefore, the VIAF IDs for all these clusters are stored locally, with the note to be manually checked regarding the accuracy.

7. Crosslinking-through author's disambiguation

Sims, Christopher A.

1. -Sims, Christopher A. Sims, Chris A. Sims, Christopher A., 1942- Christopher A. Sims Sims, C.A. (Christopher A.) Sims, Christopher - (<http://viaf.org/viaf/76452133>)

Living: 1942-10-21 -

Publications:

- 1. Advanced in econometrics / ed. by Christopher Sims. - Cambridge, 1994. - **10.54 %**
- 2. Advances in econometrics : sixth world congress - **13.61 %**
- 3. Bayesian skepticism on unit root econometrics - **16.67 %**
- 4. Business cycles : indicators and forecasting - **0 %**
- 5. Discrete actions in information-constrained tracking problems - **12.6 %**
- 6. Fiscal aspects of Central Bank independence - **100 %**
- 7. Forecasting and conditional projection using realistic prior distributions. -- - **0 %**
- 8. MCMC method for Markov mixture simultaneous-equation models a note - **100 %**
- 9. Methods for inference in large multiple-equation Markov-switching models - **100 %**
- 10. Modeling trends - **0 %**
- 11. Models and their uses - **20.41 %**
- 12. A nine variable probabilistic macroeconomic forecasting model - **16.9 %**
- 13. The precarious fiscal foundations of EMU - **33.33 %**
- 14. Rational inattention: a research agenda - **100 %**
- 15. Solving nonlinear stochastic optimization and equilibrium problems backwards - **0 %**
- 16. Toward a modern macroeconomic model usable for policy analysis - **21.08 %**
- 17. Understanding unit rooters : a helicopter tour - **18.26 %**
- 18. Were there regime switches in U.S. monetary policy? - **100 %**

Co-Authors:

- 1. Zha, Tao **100 %**
- 2. Waggoner, Daniel F. **100 %**
- 3. Sims, Christopher A **33.33 %**
- 4. Matějka, Filip **0 %**
- 5. Litterman, Robert B. **0 %**
- 6. Doan, Thomas **0 %**
- 7. Zarnowitz, Victor 1919-....) **0 %**
- 8. Leeper, Eric M. **0 %**
- 9. Leeper, Eric M. (Eric Michael) **0 %**

Sources:

- 1. BNF|13164124 - <http://catalogue.bnf.fr/ark:/12148/cb131641247>
- 2. DNB|123351022 - <http://d-nb.info/gnd/123351022>
- 3. ISNI|0000000121403165 - 0000000121403165
- 4. LC|n 87118685 - n87118685
- 5. NK|C|ola2011649974 - ola2011649974
- 6. NTA|120972670 - 120972670
- 7. NUKAT|n 2013135960 - vtls009372898
- 8. SUDOC|034696423 - 034696423
- 9. VKP|Q109737 - Q109737
- 10. NLA|000035244205 - 000035244205
- 11. NLI|004013389 - 004013389
- 12. BIBSYS|90746401 - 90746401
- 13. NII|DA08117125 - DA08117125
- 14. RERO|vtls003834321 - vtls003834321

Other links:

- <http://id.loc.gov/authorities/names/n87118685>
- <http://www.wikidata.org/entity/Q109737>
- * http://dbpedia.org/resource/Christopher_A._Sims **DBPEDIA**
- <http://data.bnf.fr/ark:/12148/cb131641247#foaf:Person>
- <http://isni.org/isni/0000000121403165>
- <http://data.bibsys.no/data/notribib/authorityentry/x90746401>
- <http://www.idref.fr/034696423/id>
- * **D-NB links - and a list of possible Publications from D-NB**
- Rational inattention: a research agenda - **100 %**
- Fiscal aspects of Central Bank independence - **100 %**

 ☒ True ☐ False

Figure 7.5 The case when the prototype has found and assessed as correct match an EconStor author with a VIAF cluster

7.4 The experimental setup

As noted earlier, for every conducted search on a particular author, the number of obtained VIAF clusters can vary from null to some hundred. The shown example is a case where a total of five clusters are retrieved as a match to that search, one is assessed correct and four others as an inaccurate match (as in Fig 7.6). However, for a given author the number of clusters retrieved as correct may also be none, one or more than one. In all cases where the prototype achieves to determine correct clusters, even when there is more than one, in a fully automatic way they are stored in a local database with the corresponding VIAF ID.

2. -Barsky, Robert B. barsky, robert - (<http://viaf.org/viaf/65064809>)

Living: 0 - 0

Publications:

1. Accounting for the black-white wealth gap : a nonparametric approach - 20 %
2. Bull and bear markets in the twentieth century - 11.79 %
3. Do flexible durable goods prices undermine sticky price models? - - 13.61 %
4. The Fisher hypothesis and the forecastability and persistence of inflation. - - 10.91 %
5. The Japanese bubble a 'heterogeneous' approach - 18.26 %
6. Measuring the cyclicalilty of real wages: how important is composition bias? - 12.31 %
7. A monetary explanation of the great stagflation of the 1970s - 21.82 %
8. Oil and the macroeconomy since the 1970s. c2004 - 0 %
9. Real wages over the business cycle - 0 %

Co-Authors:

1. Kilian, Lutz. 0 %
2. Solon, Gary 0 %
3. Kimball, Miles S. 0 %
4. House, Christopher L. 0 %
5. DeLong, J. Bradford (J.) 0 %
6. Sims, Eric R. 0 %
7. Parker, Jonathan (Jonathan A.) 23.57 %
8. Long, J. Bradford de 0 %

Sources:

1. DNB|128649585 - <http://d-nb.info/gnd/128649585>
2. ISNI|000000008251112X - 000000008251112X
3. LC|n 86005290 - n86005290
4. NTA|07482712X - 07482712X
5. PTBNP|121650 - 121650
6. NLI|002332656 - 002332656
7. BIBSYS|3073738 - 3073738

Other links:

- <http://id.loc.gov/authorities/names/n86005290>
- * D-NB links - and a list of possible Publications from D-NB
 - Why does the stock market fluctuate? - 18.26 %
- <http://isni.org/isni/000000008251112X>


 ☐ True ☐ False

Figure 7.6 The case when the prototype has depicted as incorrect a VIAF cluster

7. Crosslinking-through author's disambiguation

Williamson, Oliver E.

1. -Williamson, Oliver E. Williamson, Oliver Eaton, 1932-.... Williamson, Oliver E., 1932- Oliver E. Williamson American economist Williamson, Oliver E. (Oliver Eaton) - (<http://viaf.org/viaf/108143756>)

Living: 1932-09-27 -

Publications:

1. Antitrust economics : mergers, contracting, and strategic behavior - **16.9 %**
2. The Firm as a nexus of treaties - **0 %**
3. Defense contracts : an analysis of adaptive response - **16.9 %**
4. Economic institutions of capitalism. - **0 %**
5. Economic institutions of capitalism : firms, markets, relational contracting - **0 %**
6. Economics of discretionary behavior - **22.36 %**
7. Economics of discretionary behavior; managerial objectives in a theory of the firm - **11.95 %**
8. Ekonomiczne instytucje kapitalizmu : firmy, rynki, relacje kontraktowe - **0 %**
9. Ekonomikku oganizeshon : Torihiki kosuto paradaimu no tenkai. - **0 %**
10. Transaction cost economics - **77.46 %**
11. Gendai kigyō no soshiki kakushin to kigyō kodo. - **0 %**

Co-Authors:

1. Masten, Scott E. **0 %**
2. Winter, Sidney G. **0 %**
3. Phillips, Almarin **0 %**
4. Gustafsson, Bo **0 %**
5. Aoki, Masahiko **0 %**
6. Rubin, Paul H. **0 %**
7. 井上, 薫, 1941- **0 %**

Sources:

1. BNC|a11000880 - .a11000880
2. BNF|12253195 - <http://catalogue.bnf.fr/ark:/12148/cb12253195b>
3. DNB|124179045 - <http://d-nb.info/gnd/124179045>
4. ISNI|0000000108596900 - 0000000108596900
5. LCN|50019005 - n50019005
6. LNB|LNC10-000020200 - LNC10-000020200
7. NDL|00461060 - 00461060
8. NKC|jn20000810092 - jn20000810092

Other links:

- <http://isni.org/isni/0000000108596900>
- <http://www.wikidata.org/entity/Q232062>
- <http://libris.kb.se/resource/auth/328155>
- <http://alpha.bn.org.pl/record=a10419408>
- <http://www.idref.fr/031286941/id>
- <http://id.loc.gov/authorities/names/n50019005>

* **D-NB links - and a list of possible Publications from D-NB**

- The markets and hierarchies program of research: origins, implications, prospects : paper presented at the IIM Summer Workshop on "The Economics of Internal Organization", Berlin, June 23 - July 11, 1980 - **7.35 %**

- <http://id.ndl.go.jp/auth/entity/00461060>
- <http://data.bnf.fr/ark:/12148/cb12253195b#foaf:Person>
- * http://dbpedia.org/resource/Oliver_E._Williamson **DBPEDIA**

 ☐ True ☐ False

Figure 7.7 The case when the prototype has depicted as “maybe” a VIAF cluster

7.4.2 Storing and evaluating the prototype results

At the time we refer to VIAF to disambiguate an author from our repository, the retrieved cluster persistent identifier is stored locally, including there the prototype's decision regarding the correctness of the cluster. Beyond this, the performance of the system is evaluated by including human evaluation. In total, there are evaluated 1026 authors, and for each of them, the evaluators have assessed the system decision of the retrieved clusters. Therefore, every cluster's determined status, such as "correct", "maybe" and "incorrect", is evaluated by a user to see if it is the right decision. Figure 7.8, gives an overview of such action. As can be noted, the assessor just needs to decide if agrees (True) or disagrees (False) with the system's decision. Let's take the case (a.) from figure 7.8. In that example, the prototype has estimated that this cluster is "correct" and the user has assessed it as such. While, in the case of (b.), the user does not agree with the system's evaluation i.e. it is an incorrect cluster but the system has been depicted as correct. However, the case (f.) is the opposite example, the user picks it as the right cluster beside the fact that the system has assessed it as an incorrect match.

	prototype	user
a.	True (Green checkmark)	True (Selected)
b.	True (Green checkmark)	False (Selected)
c.	False (Yellow question mark)	True (Selected)
d.	False (Yellow question mark)	False (Selected)
e.	Incorrect (Red X)	True (Selected)
f.	Incorrect (Red X)	False (Selected)

Figure 7.8 The prototype and user evaluation for retrieved VIAF clusters

In this manner, all the cases except (e.) are stored in the database. The case (e.), which means "incorrect" by the system and by the user, does not have any relevance for further usages. Therefore, if for a selected author all the retrieved clusters belong to option (e.), then "NA" is stored instead of the VIAF ID. That means the system cannot find any match for that author in VIAF. For the options, (c.) and (d.) the user evaluates the accuracy of the cluster as correct (True) or Incorrect (False). Table 7.2 gives a concrete instance by including the prototype and user evaluations.

7. Crosslinking-through author’s disambiguation

Table 7.2 EconStor authors with the corresponding found VIAF ID

Nr	Name	VIAF ID	Prototype	User
1	Aalberts, Tanja	169218012	Correct	True
2	Abadie, Alberto	25681789	Correct	True
3	Abbassi, Puriya	171900455	Correct	True
4	Diamond, Peter A.	172333705	Correct	True
5	Diamond, Peter	172333705	Correct	True
6	Holmström, Bengt	88128957	Incorrect	False
7	Hart, Oliver	49326856	Correct	True
8	Shiller, Robert	41900524	Incorrect	False
9	Sigmund, Peter	NA	NA	//
10	Tirole, Jean	93736926	Correct	True
11	Sinn, Hans-Werner	2543709	Incorrect	True
12	Kasper, Wolfgang E.	27162644	Incorrect	False
13	Williamson, Oliver E.	108143756	Maybe	True
14	Sims, Christopher A.	76452133	Correct	True
15	Deaton, Angus	85162145	Maybe	True

The benefits of accurate and unique identification of authors within a repository / DL can serve many purposes. In the beginning, this can be used for clustering purposes, i.e., collecting author’s publications together, and facilitating the creation of a comprehensive authority profile within the DL. Beyond that, the obtained VIAF ID provides us with a consistent connection to the corresponding VIAF cluster that ensures a continuous data exchange with the possibility and to extends the range of identifiers based on the resources located there. For this purpose, in all the cases when for a given author one of the combinations (correct, true), (incorrect, false), or (maybe, true) is fulfilled, the selected cluster is depicted as accurate. As a result, the VIAF ID is formatted and stored in an RDF triple such below:

<author’s identification in EconStor> owl:sameAs <VIAF ID>

Example:

<<http://linkeddata.econstor.eu/beta/resource/authors/9133153>> owl:sameAs
<http://viaf.org/viaf/70222107>

7.5 Limitations of this approach

In a situation where an author within a repository is represented only by the name label, without having any local identifier, the disambiguation process becomes more complex. For the reason that we are forced to operate with a very small amount of metadata. In that case, we rely entirely on the record level, by considering the title, co-authors, and other information as part of the record. Therefore, the potential assignment of a discovered persistent identifier is done as part of the record, including there the record id, the name, and the assigned identifier for that author, as in the following format:

Record ID, Name Surname, Persistent Identifier

Such an approach is implemented in EconBiz where in some records the authors are identified with an identifier, such as the GND ID in that case, as in the given instance:

```
id: "10011870677", "name": "Aaberge, Rolf", "gnd_id": "170422291"
```

It is interesting to note that according to the analysis performed on EconBiz, based on the 2018 dump files, - which datasets are characterized as partially disambiguated in regard to authors - were evident 432 553 authors with GND ID (name labels with distinct GND ID). From that list, 13 817 authors share absolutely the same name (name string) but have more than one GND ID. This list precedes the name “Wang, Wei” with 36 different GND IDs, followed by “Li, Jing” with 35, “Li, Wei” with 27, “Müller, Michael” with 25, and so on. These numbers, even more, are emphasizing the importance and the role of author disambiguation in the IR and crosslinking process.

7.5 Limitations of this approach

The disambiguation process, i.e. identifying the correct author inside a repository, in this scenario is described and applied through VIAF clusters. The major experiments are performed based on EconStor content; however, the same approach is generic enough and can be applied at any other

7. Crosslinking-through author's disambiguation

repository/DL, if the necessary data is provided. The most crucial data elements are the list of publications and co-authors in both repositories. Especially, the data at the initial repository plays a very important role in this process. Therefore, one of the deficiencies that would impede the application of this approach is indeed the lack of necessary data at the initial repository.

Most of the cases in which the approach failed to identify the right author in VIAF were due to a lack of data. For example, the author "Holmström, Bengt" at the initial repository, e.g. EconStor has only one single publication and no co-author's correlations. Hence, that makes his correct identification in other repositories almost impossible, since we have nothing to compare.

Another issue that represents an obstacle in the disambiguation process is related to already assign false authorship in any of the repositories, i.e. the research output that is assigned to an incorrect author. This problem also affects the co-authorship relations and increases the complexity of matching an author in other repositories. Regarding our experiments, there was no evident example of such cases. But there were many marked cases where publications that belong to authors with the same name, are registered separately. In all these cases, the authors are differed by adding one additional initial (ex. the first letter of middle name) to one of the authors. Such an example is "Diamond, Peter" and "Diamond, Peter A.". However, as can be shown from table 7.2, the system has identified both of them with the same VIAF ID (<http://viaf.org/viaf/172333705>), as correct in both cases. This is sufficient to prove that these two strings of the name represent the same person; therefore, they can also be merged into a single name.

Part III

Evaluation and Results

Evaluation of approaches across LOD Repositories

*“A goal is a dream
with a deadline ”*

Napoleon Hill

This study presents several approaches with regard to the initial purpose to enrich scientific publications of a DL with other relevant information from other repositories. As relevant information can be considered a list of closely related publications stored and indexed in other repositories, even if they belong to different domains. Therefore, the main challenge relies on the crosslinking and retrieving process, i.e. the determination of semantic relatedness between the initial and retrieved publications. Starting from the aligned concepts between the LOD repositories, we extended our research into two additional approaches for measuring and determining semantic relatedness. The main part of this chapter is previously published in a journal article [HaTo17] and research papers [HaLT14, HaTo16, RaHL16].

8.1 Results and Discussions

As emphasized in chapter 6, the use of alignments between LOD repositories is a productive step for retrieving an initial set of publications from targeted repositories. Especially, alignments are useful for reformulating a search query from one vocabulary to another [BiTu16, HaLT14, JJHY12]. At the same time, the presence of thesauri descriptors at initial or targeted repositories has a huge impact on metadata enrichment.

8. Evaluation of approaches across LOD Repositories

The previously generated results showed that relying only on the aligned concepts between repositories/thesauri the list of retrieved results is very wide. Therefore, further processing steps are necessary to narrow this subset and generate the relevance-based ranking. According to this, the implementation of data mining approaches was considered mandatory. In total two main approaches have been evaluated, for measuring the semantic similarity between the initial publication with the retrieved publications as a result of these alignments.

Therefore, the implementation of the count-based approach through TF-IDF and Cosine Similarity requires a large set of publication's metadata, to measure and generate a similarity degree. Moreover, the right combination of metadata elements is crucial. Hence, in several cases, the frequency of a more general concept in these metadata had a negative impact on the result. For example, regarding the publication titled *"Food prices and political instability"*, the word *"food"* has been determinant in the similarity measurements. Thus, the retrieved publications have been related to *"agriculture"*, *"food security"* or *"health"* rather than *"food prices"* or *"politics"*, which semantically are not close to the initial publication. Different adjustments among the metadata components are resulting in improvements considering the retrieved results. However, this applies heuristic involvements in the evaluation of results. Moreover, the count-based approach shows significant weakness in recognizing relationships among terms, even in cases when the presence of thesauri is evident. Therefore, its performance is strictly related to the presence of the same words among the compared texts. In order to overcome such limitations, we have investigated word embedding, as the most comprehensive and promising approach. The evaluations are done comparatively, in both approaches at the same time, on selected repositories. The generated results of top-ten retrieved publications are assessed through human judgments regarding their relevance to the triggered publication.

Nowadays, there are several research articles that at the center have the evaluation of word relatedness i.e. semantic similarity among words, based on the word embedding approach. Almost, all of these evaluations take place in already human-annotated datasets such as WordSim353 [FGMR01] or SimLex-999 dataset [HiRK16]. Another set of publications are focused on IR, by evaluating the binomial query - retrieved documents, or question -

answer. Even in these cases, there are present several humanly annotated datasets, such as TREC [HCRP07] or PubMed [LiWi07], with already predefined thresholds. However, even our case represents a common IR task, we find it more appropriate for evaluating the proposed approaches on tangible crossdomain repositories.

The main task in our case relies on semantic relatedness among documents, i.e. publications from different domain repositories. Therefore, there is an obvious difference in how the retrieval is initiated. We are starting by considering all the metadata behind a publication, rather than a user-entered query. When a user makes a query, it is consisted of carefully chosen appropriate terms, without “noisy” words in it. While at the publications metadata, the importance of metadata components i.e. title, abstract, keywords, should be determined additionally. Except that, the weight of the words inside these components plays a crucial role. Thus, different combinations among these metadata result in different retrieved publications. This is one of the reasons, for performing our evaluations on these types of datasets.

8.1.1 The results

As mentioned before, in total 57 publications are evaluated, including different sets of metadata with the two applied approaches. The process is described in detail in section 6.2, regarding VSM, and section 6.3 concerning the WE approach. Figure 6.6 and figure 6.7 give more details about them. For each of these 57 EconStor publications, the prototype has retrieved 300 publications from the targeted repository i.e. AGRIS. Iteratively we have evaluated the top-ten retrieved publications, ordered on both approaches, with two different sets of metadata (all the metadata versus titles). Thus, for each of these EconStor publications p_i , a set of publications D_i is retrieved, where $D = \{d_1, d_2, d_3, \dots, d_{300}\}$ is a subset of AGRIS repository.

Table 8.1 depicts an example of two such evaluations. By default as a reference is taken the ordering done on Cosine Similarity score, denoted as *top10CS*. After that, for each EconStor publication i.e. *publication1*, *publication2*, the retrieved results are ordered by Word Embedding approach, similarity score, denoted as *topW2V*. Therefore, the relevance of

8. Evaluation of approaches across LOD Repositories

the retrieved publications is judged and labeled manually with **i** - *irrelevant*, **s** - *somehow* and **r** - *relevant*.

For clarifying this, let us have a closer view of table 8.1. Considering the EconStor *publication 1*, the first retrieved result based on Cosine Similarity is evaluated as irrelevant (i), while the first ranked result based on Word2Vec is judged as relevant (r). Thus, at the end of each column, cumulatively are shown the evaluation results of both approaches, concerning the relevance. The generated results make evident the discrepancies between the applied approaches.

Table 8.1 An example of the top-ten retrieved and evaluated publications for two EconStor publications, ordered in both approaches with different sets of metadata

publication 1 (p ₁)								publication2 (p ₂)																
all metadata				only titles				all metadata				only titles												
rank	top10CS		relCS	topW2V		relW2V	titleCS		relTcs	titleW2V		relTw2v	top10CS		relCS	topW2V		relW2V	titleCS		relTcs	titleW2V		relTw2v
1	d ₁	i	d ₅	r	d ₈	r	d ₂₂	s	d ₁	r	d ₁	r	d ₈	s	d ₁₂	s								
2	d ₂	r	d ₅₉	r	d ₁	i	d ₂	r	d ₂	s	d ₃₃	r	d ₇	s	d ₈	s								
3	d ₃	r	d ₂₈	i	d ₆	i	d ₃	r	d ₃	r	d ₁₃	r	d ₂₃	i	d ₅	s								
4	d ₄	i	d ₅₇	s	d ₄	i	d ₇	s	d ₄	s	d ₄₁	s	d ₁₂	s	d ₂₃	i								
5	d ₅	r	d ₃₉	s	d ₅	r	d ₆	i	d ₅	s	d ₂₇	s	d ₄	s	d ₁₉	i								
6	d ₆	i	d ₆₀	i	d ₂	r	d ₅₉	r	d ₆	r	d ₃	r	d ₁₄	s	d ₁	r								
7	d ₇	s	d ₄₂	i	d ₁₃	i	d ₁₄	r	d ₇	s	d ₅	s	d ₁₀	s	d ₂₈	i								
8	d ₈	r	d ₃₄	s	d ₂₃	i	d ₄₂	i	d ₈	s	d ₂₀	r	d ₁	r	d ₇	s								
9	d ₉	i	d ₆₆	i	d ₄₆	i	d ₃₉	s	d ₉	s	d ₃₆	s	d ₆	r	d ₂₂	s								
10	d ₁₀	i	d ₃	r	d ₃	r	d ₆₀	i	d ₁₀	s	d ₁₅	r	d ₁₇	s	d ₂₇	s								
r	4		3		4		4		3		6		2		1									
s	1		3		0		3		7		4		7		6									
i	5		4		6		3		0		0		1		3									

Accordingly, Word2Vec ranks, referring again to *publication 1* (p_1) in table 8.1, the 59th retrieved publication according to CS (d_{59}), as fifth (d_5). At the same time, there are evident several cases when Word2Vec has re-ranked in top-ten publications that CS has ordered below 100.

Regarding the top-ten retrieved publications, based on all metadata, the Word Embedding approach gives 70.9% completely different list of documents in top-ten, versus Vector Space Model. Thus, only 29.1% of the same retrieved publications appear in top-ten, by both approaches. These cases are shown in table 8.1 with a highlighted background. On the entire set of evaluations, with all metadata, the Vector Space Model i.e. TF-IDF with Cosine Similarity gives 16.4% relevant publications in top-ten, 42.7% somehow relevant and 40.9% irrelevant. While on the same set, Word Embedding i.e. Word2Vec gives 17.3% relevant publications, 42.7% somehow relevant and 40% irrelevant. A better graphical representation of these data is visible in figure 8.1.

At first glance, it seems a minor difference in generated results between both approaches, according to the relevance of the top-ten retrieved publications. However, a more detailed analysis shows quite interesting occurrences, highlighting the differences and similarities between them, as presented below.

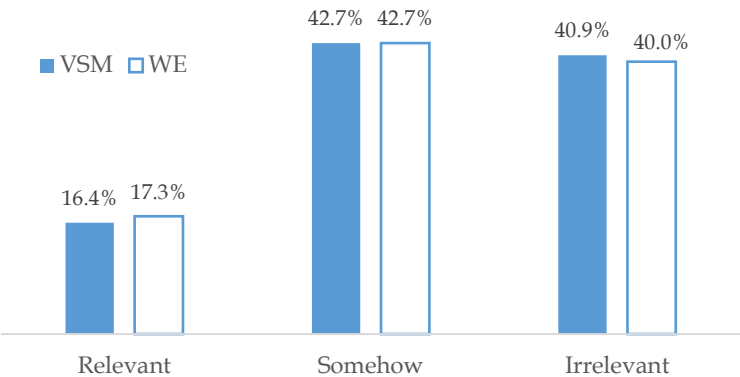


Figure 8.1 Humanly evaluation of top-ten retrieved publications based on the Vector Space model and Word Embedding approach.

8. Evaluation of approaches across LOD Repositories

If the analyses are distributed only to the list of relevant publications in top-ten, namely by excluding common relevant publications, i.e., the disjunctive union $rel_{CS} \Delta rel_{W2V}$, then W2V catches 27% of all relevant publications, while CS with TF-IDF 13.5%. Thus, both of them perform with 40.5% difference (the value of disjunctive union), or 59.5% in the same fashion, according to the number of relevant publications in top-ten. Concerning the irrelevant documents, WE gives 40.4% versus 46.1% of VSM, in that list. Figure 8.2 highlights more details about these proportions.

We also note that WE is able to generate better results, as far as relevance is concerned. It also reaches to ‘seize’ publications that even have little or no similar concepts among themselves. This is because of WE’s ability to present correlations between words.

The number of irrelevant results is in the frame of expectations, taking into account the different domains between the repositories where the evaluations take place. In the case when the selected publications are purely economic, such as *“Taxes, wages and working hours”*, both approaches give zero relevant recommendations, and four somehow relevant. Conversely, for inter-domain publications such as *“Politics, globalization, and food crisis discourse”*, or *“Public policies against global warming”* the system achieves to retrieve four very relevant publications. The other reason is related to the limited number of records for each search at the target repository. For evaluation purposes, the prototype processes only 300 publications, for every EconStor paper at that repository, i.e. AGRIS. Increasing that number means increasing the possibility for more relevant publications, but at the same time increasing the cost of processing.

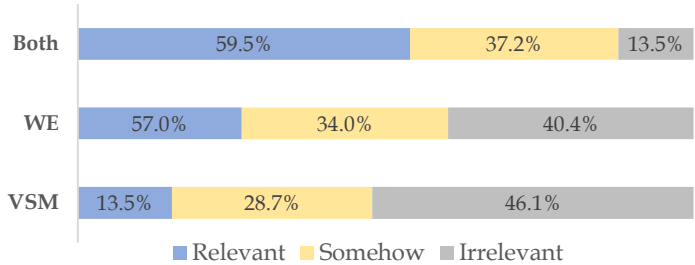


Figure 8.2 The relevance of the retrieved result based on the Vector Space Model and Word Embedding approaches, separately by including the common results.

In addition to the given metadata, Word embedding achieves good performance in smaller texts also [GMSB17, KeRi15]. Simultaneously, we have analyzed and evaluated the relevance of the retrieved documents when ordering score is used only the similarity between titles. For example, between the titles “Do inflation and high taxes increase bank leverage?” and “Are government regulations pushing food prices higher?” the Word2Vec has scored 0.7223 similarity degree, versus zero to CS score.

The results presented in Figure 8.3 point out the slight domination of WE in terms of performance only in titles. Therefore, WE achieved to retrieve 12.7% relevant publications versus 11.0% of VSM. In addition, VSM retrieves 8.8% more irrelevant documents than WE.

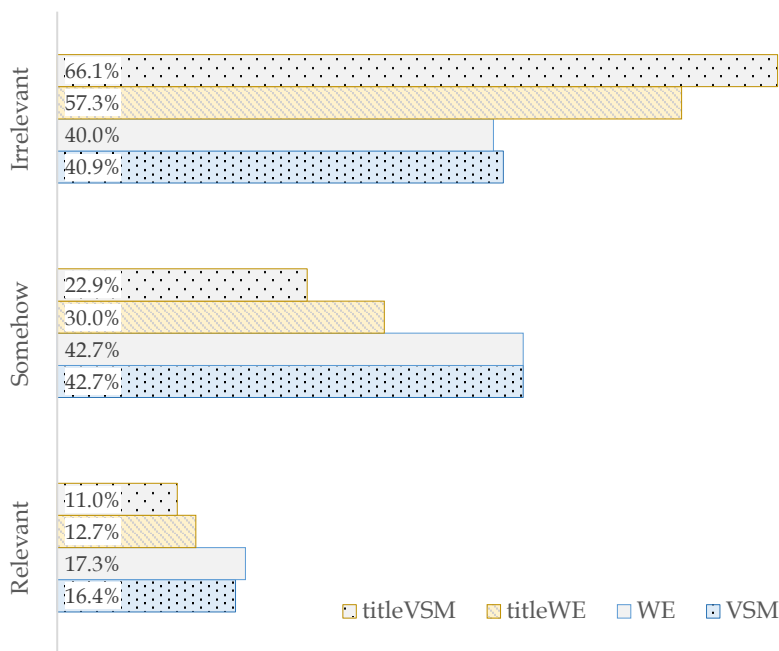


Figure 8.3 The relevance of the retrieved result based on VSM and WE approaches, generated in all metadata versus titles.

8. Evaluation of approaches across LOD Repositories

The score generated as the combination of all the metadata, i.e. $sim[(2p_t, p_{abs}, K_p, D_p), (2d_t, d_{abs}, D_d)]$ achieves to catch 17.3% more relevant or somehow relevant publications rather than the score calculated on titles, $sim(p_t, d_t)$, referring to the Word Embeddings approach. Furthermore, only 28.2% of publications in top-ten are the same in both ordering scores.

8.1.2 Cumulative Gain measures

A formal way for presenting the results is done by applying the Discounted Cumulative Gain (DCG) measure, as the most notable metric for quantifying the performance of ranking high relevant documents [JäKe00]. The formulation of DCG is defined in section 4.3, through equation 4.4. The number of evaluated documents in our case is continually 10.

The application of DCG in our evaluated data requires translation of the relevance values from literals to numbers. Such that, **r** that stands for relevant is denoted with **2**, **s** of somehow's as **1**, and **i** for irrelevant as **0**. Thus, in total there are three relevance values, $rel_i \in \{0,1,2\}$. Table 8.2 embodies exactly the *publication 1* from table 8.1, after including these translations. As can be noted in table 8.2, the DCG score is calculated for both approaches, i.e. CS and W2V on all metadata and titles comparatively. Therefore, the four ranking strategies are shown. The end of each column gives the sum of these values as stated in the formulation. Therefore, considering the same example, the DCG_{10} score for Cosine Similarity on all metadata is 4.0 while the DCG_{10} score of Word2Vec on the same metadata is 4.973.

However, the DCG score is not the best solution for measuring the performance of several approaches with different sets of metadata, regarding the ranking of relevant documents [JäKe02, WWLH13]. For that purpose, several other modifications of DCG are present for different circumstances. In our case, since we are operating with the fixed number of evaluated documents over all the approaches, i.e. ten, the normalized discounted cumulative gain (nDCG) is applied. For this purpose, the normalization between the results, based on the relevance order is performed. For each of the columns in table 8.2 ($relCS$, $relW2V$, $relTcs$,

$relTw2v$), the DCG is recalculated after sorting the retrieved documents in decreasing order of relevance. For example, $relCS$ now will be ordered such as $(2_12_22_32_41_50_60_70_80_90_{10})$. The DCG value calculated in this way is known as the Ideal DCG (IDCG). Hence, $relCS$ will have $IDCG_{10}$ of 5.510. The normalized discounted cumulative gain (nDCG), represents the fraction of DCG with ideal DCG. In this case, for the $relCS$ example in table 8.2, we have $nDCG_{10}=4.0/5.51 = 0.726$.

Table 8.2 An example of generating DCG_{10} score on top-ten retrieved publications for one EconStor publication.

<i>publication 1 (p_1)</i>								
	all metadata		only titles		all metadata		only titles	
position (i)	$relCS$	$relW2V$	$relTcs$	$relTw2v$	DCG CS	DCG W2V	DCG titleCS	DCG titleW2V
1	0	2	2	1	0.000	2.000	2.000	1.000
2	2	2	0	2	1.262	1.262	0.000	1.262
3	2	0	0	2	1.000	0.000	0.000	1.000
4	0	1	0	1	0.000	0.431	0.000	0.431
5	2	1	2	0	0.774	0.387	0.774	0.000
6	0	0	2	2	0.000	0.000	0.712	0.712
7	1	0	0	2	0.333	0.000	0.000	0.667
8	2	1	0	0	0.631	0.315	0.000	0.000
9	0	0	0	1	0.000	0.000	0.000	0.301
10	0	2	2	0	0.000	0.578	0.578	0.000
DCG ₁₀					4.000	4.973	4.064	5.373
IDCG ₁₀					5.510	5.436	5.123	6.200
nDCG ₁₀					0.726	0.828	0.793	0.867

8. Evaluation of approaches across LOD Repositories

The interpretation of scores can lead us to better understanding the performance of proposed approaches. The computed DCG_{10} and $nDCG_{10}$ scores are visualized in figure 8.4 based on the 57 evaluated EconStor publications. Therefore, from the same figure can be noted that DCG value shows a better performance of W2V versus CS in both metadata sets. When all the metadata are considered, the DCG_{10} of W2V is 4.057 while CS is 3.861. This insight specifies that W2V archives to show in top-ten much relevant documents than CS. The discrepancy is even more notable when only titles are considered, i.e. 3.069 versus 2.317 in favor of W2V.

However, an interesting sighting shows the analysis of $nDCG_{10}$ score. The value of 0.869 at CS comparing to 0.835 at W2V let to know that CS achieve to perform a better ranking of the relevant document. Thus, although W2V attains to caught more relevant or somehow relevant documents in top-ten, CS achieves to perform better ranking. Nonetheless, it is not a case when the comparison is done only on titles. Emphasizing, even more, the dominance of W2V in short texts.

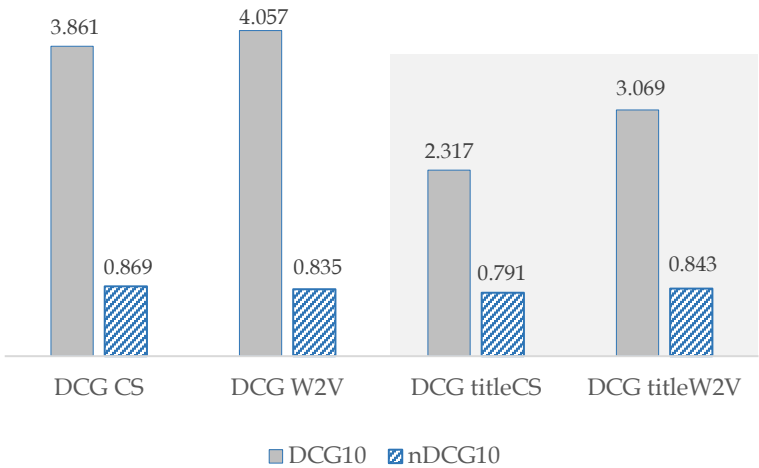


Figure 8.4 The average Discounted Cumulative Gain (DCG) and Normalized DCG (nDCG) score for VSM and WE approaches, generated on two different metadata sets.

8.2 Summary

This chapter puts the focus on the advantages resulting from improved interoperability among different Digital Libraries by evaluating different algorithms to achieve this interoperability. For this purpose, bibliographic Linked Open Data repositories are considered by investigating the alignments among them. The evaluated results show that the list of retrieved publications according to each aligned concept between repositories was extremely wide. While the attempt to find publications in the target repository, with the same set of descriptors as in the initial one, results in no publications returned. Therefore, we use alignments between repositories for retrieving an initial set of publications, especially as an important component for determining the weight of the terms in the metadata set.

The semantic relatedness of the retrieved publications with the triggered publication is measured by applying two main approaches comparatively. The generated results show that the traditional count-based and text-matching approach through TF-IDF and Cosine Similarity, are satisfactory. However, it relies on heuristics to determine a higher level of semantic similarity among publications. Its performance is closely related to the common words among the compared publications. The disability for determining the word relatedness appears to be the main weakness, even in the cases when the presence of thesauri is evident.

Given this, we followed the deep learning approach to model semantic word representations. The implementation of contemporary Word2Vec results in important outcomes. This is achieved by simplifying the combination process between the metadata, and even more, by performing it on a smaller set of metadata, such as title's concepts only. Substantial improvements are evident by extending the set of metadata with concepts from the abstract and keywords. The results show that the implementation of the word embedding approach achieved to retrieve as the top-ranked relevant recommended publications, which the previous approach has ranked far below from the top positions. Therefore, 27% of all relevant publications are caught by Word2Vec only, while 13.5% by CS with TF-IDF. Thus, they are performing with a 40.5% difference concerning the outcome

8. Evaluation of approaches across LOD Repositories

of relevant retrieved publications. A proper interlacement between these approaches brings to promising improvements.

In addition, the results are presented by applying the Discounted Cumulative Gain (DCG) measure and Normalized Discounted Gain (nDCG). These scores prove a light dominance of Word2Vec to show in top-ten much relevant documents than CS. The discrepancy is even more notable when only titles are considered, regarding the DCG₁₀ score of 3.069 for Word2Vec, versus 2.317 for CS. However, although W2V attains to caught more relevant or somehow relevant documents in top-ten, the nDCG₁₀ value indicates that CS achieves to perform better ranking when performing on all the metadata set.

In conclusion, as a result of the applied approaches, publications stored in a particular repository, i.e. digital library are enriched with closely related semantic recommendations from other Linked Open Data repositories. This will enhance the visibility of publications from a single place by sparing the scholar for further navigation in other digital libraries. The research can be extended with several other combinations of the proposed approaches and metadata. At the same time, new methods can be introduced. However, in any case, a human judgment regarding the relevance of retrieved results is necessary. Meantime, these judgments know to be expensive and inconsistent.

Evaluation of author's disambiguation

*“Linking up the things you were
with the things you become
is what growing up is”*

James L. Brooks

This chapter describes the results of the approaches regarding author's centered metadata, presented in Chapter 7, and the benefits of author persistent identifiers. The evaluations are done through the developed prototype, used for the assessment of the proposed algorithms. Essentially, the prototype acts on two levels. Initially, it automatically checks VIAF for a particular author and automatically determines the appropriate clusters according to the principles presented in chapter 7. For each found cluster, the VIAF ID is assigned to the corresponding author in the initial repository (EconStor in our case). The next level is related to the extraction of the results found in the cluster, and by redirecting several queries to the corresponding digital libraries that the cluster aggregate. As a result, the author's profile is enriched with additional relevant information, which previously has not been part of that repository. Parts of this chapter are published in several peer-review publications [HaPT21, HaRT15, PiHa21].

Moreover, the assigned identifiers are extended with several others by considering WIKIDATA as a linking hub, see figure 9.6. The role of the community, in this case, is undisputed in terms of contributions to the data population. Consequently, linking more identifiers offers the opportunity to find and collect various other related information to a particular author in a single place.

9. Evaluation of author's disambiguation

9.1 Evaluation of VIAF approach

Chapter 7, especially section 7.1 underlines the repository disambiguation level that highlights which authors need to be identified and clustered, plays an important role in the overall process. That's because a partially or entirely disambiguated repository/DL can ensure a larger and precise list of publications and co-authors network to a particular author. In addition, the presence of a persistent identifier facilitates the generation of expanded author profiles enriched with other information about her/his.

In principle, the disambiguation process must rely on a record-based approach, i.e. each bibliographic record should be checked separately for each author present in that record. Therefore, in all cases when authors are not assigned with any kind of persistent identifier and their disambiguation is impossible to be done from the data within the repository, an external source should be used for that purpose. Hence, the use of VIAF is seen as a highly acceptable and efficient choice.

Finding the right VIAF authority cluster that matches the author's name presents a fundamental challenge. The automated process for checking, disseminating and selecting the best match does not always turn out to be straightforward. Therefore, as described in chapter 7, the approach is evaluated manually from individuals, on 1026 randomly selected authors from the EconStor repository. As result, the evaluation metrics of recall, precision, and F1 score are generated.

In these cases, the precision is considered for the clusters that are retrieved as correct match, i.e., all the VIAF clusters which the system has assessed as correct for representing a particular author. Referring to figure 7.8, these are the cases under the option (a.) and (b.). In fact, it is the fraction among the really correct clusters (true positive) with all the clusters that the system has presented as correct, including all clusters that are assessed as correct, but users have not agreed to this (false negative). This form of evaluation tends to bring the approach into action, to measure the ability of the system to act independently of user intervention. Regarding this, the "maybe" options are also included in this calculation.

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

9.1 Evaluation of VIAF approach

The recall is the fraction between the really correct clusters (evaluated by the system and users as such) with all the clusters depicted by the system as correct, including the clusters assessed as incorrect, but the users have identified them as correct (false negative). In figure 7.8, as truly correct are results under option (a.) while “all correct” means option (a.) and (f.).

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

As mentioned in chapter 7, the performance of the system is evaluated manually, regarding the accuracy of matches. Thus, except the recall and precision, the F1 score is calculated additionally. Since for a given author the number of retrieved clusters can vary from zero to many, each of these cases is analyzed in particular.

$$\text{F1score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

The results in table 9.1 represent the clusters that are marked as positive, i.e., the prototype has marked them as correct clusters, and each of them is evaluated by the user. As can be noted, the analyses are covering the authors for which the prototype has found one, two, three and more than three clusters. While, for 265 authors (25.83%) of the evaluated instances, the system has not retrieved any match. It is worth noting that these data might change almost whenever we search through the respective clusters, taking into account the continuous updates of VIAF clusters.

Table 9.1 The number of found VIAF clusters for EconStor authors.

Number of checked authors from EconStor	Number of found clusters in VIAF	%	Precision	Recall	F1
618	1	60.23%	0.994	0.976	0.985
96	2	9.36%	0.957	0.968	0.963
40	3	3.90%	0.952	0.912	0.931
7	> 3	0.68%	0.951	0.833	0.888
265	0	25.83%	/	/	/

9. Evaluation of author’s disambiguation

What is easily noticeable, the F1, recall, and precision scores are distributed entirely in different fashion according to the number of found clusters. Hence, when for a given author only one VIAF cluster is found, the precision is 0.994, and the recall is 0.976. Therefore, the calculated F1 score, in this case, is 0.985. In all these cases, when the prototype achieves to provide only one cluster as correct, the probability of having the correct match, is almost maximal. Otherwise, when two clusters are shown as correct, the possibility of both of them being correct is not so absolute. Hence, in that situation, the precision is 0.957 with 0.968 recall, while 0.963 is the F1 value. However, in all the cases when more than one cluster is identified as correct, the different distribution of data inside them is an indicator to evaluate and select the best option. Figure 9.1 gives a better visualization of these data, by highlighting the F1 score for all the cases.

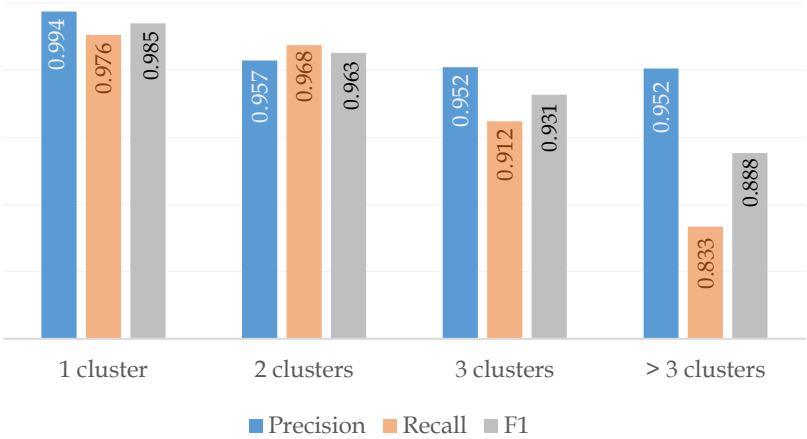


Figure 9.1 The evaluation results based on the accuracy of the found clusters.

In overall, based on the total number of evaluations (761 with at least one cluster), the efficiency of the approach is measured at 0.975 as F1 score. The precision and recall, in that case, are 0.987 and 0.963 respectively. As a result of the constant updates of the VIAF clusters, at different times we may have different results, maybe even better.

9.2 The outcome after identification

The right identifier, such as the VIAF ID, ensures us a direct retrieval of all relevant information found on that cluster and beyond. Hence, the approach makes it possible to be achieved an author profile enrichment at the initial repository. Such enrichment may include additional name variations of that author, an extended list of publications, new co-authorship correlations, and other biographic information.

For having a closer view, let us consider a specific author name from our repository, through the usage of the developed prototype. By selecting the Nobel Prize winner, “**Sims, Christopher A**”, the EconStor repository is currently showing nine publications and a list of three co-authors (see Fig. 9.8). The application of the presented approach to identify and match the corresponding VIAF cluster is resulting in significant profile data enrichments. Hence, except for the usage of the identified VIAF ID, the profile of this author has been also extended with several other information as discussed in the following section.

9.2.1 Using the VIAF cluster

Initially, we are able to retrieve a list of name variations, as alternatives to how this author appears in other digital libraries that consisting this VIAF cluster. Such output can be shown in figure 9.2, by including other biographical details, information that was missed in our initial repository. However, by expanding the list of identifiers, we are able to target other sources, such as the GND authority files or Wikipedia, to discover more information of this nature.

Sims, Christopher A.	
Sims, Chris A.	
Sims, Christopher A., 1942-	
Christopher A. Sims	
Sims, C.A. (Christopher A.)	
Sims, Christopher	
Living: 1942-10-21-	

Figure 9.2 Name variations for a particular author and living year(s).

9. Evaluation of author's disambiguation

As part of the cluster, a wider list of publications and co-authors can be found. The list of publications that this author has in that cluster is showed in figure 9.3. Thus, by comparing the publications in the initial repository (fig.7.2) with the publications found inside the VIAF cluster, there is noted an evident distinction. Hence, the list of publications in the initial repository may be enriched with additional 13 new titles.

Publications:

1. [Advanced in econometrics / ed. by Christopher Sims. - Cambridge, 1994.](#)
2. [Advances in econometrics : sixth world congress](#)
3. [Bayesian skepticism on unit root econometrics](#)
4. [Business cycles : indicators and forecasting](#)
5. [Discrete actions in information-constrained tracking problems](#)
6. [Fiscal aspects of Central Bank independence](#)
7. [Forecasting and conditional projection using realistic prior distributions. -](#)
8. [MCMC method for Markov mixture simultaneous-equation models a note](#)
9. [Methods for inference in large multiple-equation Markov-switching models](#)
10. [Modeling trends](#)
11. [Models and their uses](#)
12. [A nine variable probabilistic macroeconomic forecasting model](#)
13. [The precarious fiscal foundations of EMU](#)
14. [Rational inattention: a research agenda](#)
15. [Solving nonlinear stochastic optimization and equilibrium problems backwards](#)
16. [Toward a modern macroeconomic model usable for policy analysis](#)
17. [Understanding unit rooters : a helicopter tour](#)
18. [Were there regime switches in U.S. monetary policy?](#)

Figure 9.3 The list of publications inside a VIAF cluster for a particular author.

The same author at the initial repository has only three co-authors (fig 7.2), while the corresponding VIAF cluster is extending that list by additionally five new co-authors. Figure 9.4 shows the list of co-authors in the cluster, where the first two authors already are present in the initial repository, while the 7th and 8th records may represent the same person.

9.2 The outcome after identification

Co-Authors:

1. Zha, Tao
2. Waggoner, Daniel F.
3. Matějka, Filip
4. Litterman, Robert B.
5. Doan, Thomas
6. Zarnowitz, Victor 1919-.....)
7. Leeper, Eric M.
8. Leeper, Eric M. (Eric Michael)

Figure 9.4 The list of co-authors inside a VIAF cluster for a particular author.

Furthermore, each cluster contains identifications of libraries or institutions that contribute to the content, which appear under the `<nsl:sources>` tag, if we refer to the XML version of the cluster. Visualization of this data provides a result as in Figure 9.5. We are considering these IDs as very valuable information for expanding and enriching an author profile with new information that may not be part of the current cluster. Therefore, by having such an ID, e.g., 123351022 for DNB or n87118685 for LC, we can search directly in these repositories to link and get exactly the data about that author.

Sources:

1. BNF|13164124 - <http://catalogue.bnf.fr/ark:/12148/cb131641247>
2. DNB|123351022 - <http://d-nb.info/gnd/123351022>
3. ISNI|0000000121403165 - 0000000121403165
4. LC|n 87118685 - n87118685
5. NKC|ola2011649974 - ola2011649974
6. NTA|120972670 - 120972670
7. NUKAT|n 2013135960 - vtls009372898
8. SUDOC|034696423 - 034696423
9. WKPI|Q109737 - Q109737
10. NLA|000035244205 - 000035244205
11. NLI|004013389 - 004013389
12. BIBSYS|90746401 - 90746401

Figure 9.5 The list of sources in a particular VIAF cluster

9. Evaluation of author's disambiguation

9.2.2 Using cluster's sources

The use of these resources is done by consuming the provided APIs and similar web services. Besides, a very practical way of accessing the data is by querying the available SPARQL endpoints or local deployment of the linked data repositories. However, this way does not always ensure the most up-to-date information. Therefore, most libraries, such as DNB, LoC, BNB, in addition to providing their data or metadata in the form of dump files, provide various web services for accessing up-to-date resources. For instance, the following web service will provide publications from DNB, by adding the ID of the source for a given author found inside a cluster. Hence, the DNB ID (i.e., the GND ID) for this author is 123351022 (see fig 9.5).

D-NB URL service

<https://portal.dnb.de/opac.htm?method=simpleSearch&query={DNBauthorid}>

The output of this query, respectively the Atom²⁰ XML-based file format, is resulting in two publications, which already are part of the VIAF cluster (also highlighted in figure 7.5), such as:

1. *Rational inattention: a research agenda*
2. *Fiscal aspects of Central Bank independence.*

However, in some other cases, the exploration of these sources may reveal publications that are not yet part of the cluster. Hence, in addition to expanding the list of publications, this increases the possibility for a more accurate assessment of the cluster (depending on the update frequency).

Library of Congress (LoC) is another example of offering a linked data service for retrieving resources. A possible way for its usage can be through the token when an exact match is targeted. The token, in this case, would be *n87118685*, by considering the same author. Hence, the LoC linked data service can be explored in a way as below:

<http://id.loc.gov/search/?q=token:{LCauthorid}>

The results of such a query furthermore can be processed by examining one of the provided formats, such as JSON, RDF, etc.

²⁰ <https://portal.dnb.de/opac.atom?method=search¤tResultId=auRef%3D123351022%26any>

9.2 The outcome after identification

In addition, the Library of Congress and the German National Library offer further refined approaches to access the data, with many other possibilities. Therefore, though the SRU (Search/Retrieve via URL), as a standardized web service protocol for querying databases on the internet, makes their catalog available to everyone. Normally, for requesting the data, there is a need for previous registration and authorizations.

The DNB provides access to bibliographic and authority data by consuming the linked open data dump files. Such that, for a given entity i.e. person, using the GND ID we can retrieve a large amount of data to fill out the profile. Similar information may be harvested through the DNB's data service known as "Entity Facts"²¹, which provides information on entities of the GND Authority File.

9.2.3 Further data enrichment

Moreover, in numerous cases, a VIAF author's cluster offers mappings to several other sources, including DBpedia and WIKIDATA. We consider this as an opportunity to extend the profile of the author with several non-library resources. The prototype automatically executes a SPARQL query in DBpedia or WIKIDATA and retrieves information such as a short biography, an author picture, a link to Wikipedia page, a downloadable list of works, and other links to different sources, if there are available. The links to other sources are treated as a very valuable bit of information, as in this way the linked data graph of this author is expanded for the purpose to harvest as much as possible data.

Below is an example of getting a list of identifiers for a given author, based on WIKIDATA ID. Such that, the Google Scholar, ORCID, RePEc, ISNI, SSRN, Nobel ID, and Twitter may be retrieved for this author, by querying WIKIDATA SPARQL Endpoint (<https://query.wikidata.org>) with the source ID found inside the cluster (see fig. 9.5, **WKP** | Q109737). As result, the query from the listing 9.1 will provide the Google Scholar ID "*uXNOHdAAAAA*", short RePEc "*psi12*" and the Nobel ID "*2011/sims*".

²¹ https://www.dnb.de/EN/Professionell/Metadatendienste/Datenbezug/Entity-Facts/entityFacts_node.html, accessed 30.11.2019

9. Evaluation of author's disambiguation

Listing 9.1 Getting other identifiers from WIKIDATA

```
SELECT distinct ?gsid ?orcid ?repec ?isni ?nobelid ?ssrn ?tw WHERE {  
  OPTIONAL {wd:WKP wdt:P1960 ?gsid}.  
  OPTIONAL {wd:WKP wdt:P496 ?orcid}.  
  OPTIONAL {wd:WKP wdt:P2428 ?repec}.  
  OPTIONAL {wd:WKP wdt:P213 ?isni}.  
  OPTIONAL {wd:WKP wdt:P3188 ?nobelid}.  
  OPTIONAL {wd:WKP wdt:P3747 ?ssrn}.  
  OPTIONAL {wd:WKP wdt:P2002 ?tw}.  
}
```

In addition to this, the same source can provide a different range of data, for this author, such as short biographical data, country of origin, publications, award received, etc. Such an example is the listing 9.2., which provides all the economic awards (Q17701409) for a given author (WKP). The advantages of this approach also take account of the ability to update the data in an easy and quick manner, by anyone who will contribute.

Listing 9.2 Economic awards for a particular author based on WIKIDATA

```
SELECT DISTINCT ?label WHERE {  
  wd:WKP p:P166 ?statement.  
  ?statement ps:P166 ?award.  
  ?award wdt:P31 wd:Q17701409.  
  ?award rdfs:label ?name.  
  OPTIONAL { ?statement pq:P585 ?date. BIND(YEAR(?date) AS ?year) }  
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }  
  BIND(CONCAT(STR(?year), " - ", str(?name), "" ) AS ?label ).  
  FILTER (LANG(?name) = "en").  
} ORDER BY ASC(?awardLabel) ASC(?year)
```

In all the cases when the matched VIAF cluster is aligned to DBpedia, the profile may be enriched with several other details. Even when the VIAF does not contain such identifier, other HUBs, such as GND or WIKIDATA are considered for further expansions (see Fig. 9.6). The prototype automatically checks for the existence of such links, and in case of presence, the DBpedia SPARQL endpoint is queried. Such a query is showed under the listing 9.3.

Listing 9.3 Getting author’s information from DBpedia.

```
SELECT distinct ?abs ?birth ?picture ?link
WHERE {
  <author_id> dbo:abstract ?abs.
  OPTIONAL {<author_id> dbo:thumbnail ?picture}.
  OPTIONAL {<author_id> dbo:birthDate ?birth}.
  OPTIONAL {<author_id> dbo:wikiPageExternalLink ?link}.
  FILTER (langMatches(lang(?abs), "en")).
}
```

As we have mentioned on many occasions, WIKIDATA represents a comprehensive hub of linking data including authorities. Hence, by discovering any of the globally known identifiers, let ‘say GND ID, gives the possibility to access several others. Figure 9.6 gives an overview of some identifiers supported in WIKIDATA.

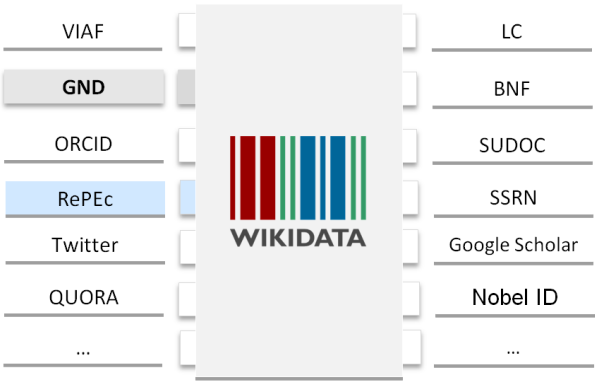


Figure 9.6 WIKIDATA as authority linking hub

Furthermore, it is very significant that the community can contribute to those data by completing the list of identifiers. For example, if an author is assigned only with the RePEc ID, once the GND ID is discovered, the same can be registered in that profile. This can be very useful in case we want to adopt a single identifier for uniquely identifying authors in a repository. Let us give some details about this. Assuming that in a particular repository

9. Evaluation of author’s disambiguation

we have several authors identified with ORCiD, some others with RePEc, VIAF, and GND ID. However, by creating an author profile, we are interested in having a unified way of representations, considering a unique identifier. Therefore, instead of creating a local hash table for mapping these ids, the usage of WIKIDATA may be considered.

At the moment WIKIDATA has impressive coverage of people, more than 8 million humans (Q5), and several other identifiers are mapped there. Figure 9.7 shows the presence of some IDs inside WIKIDATA, by visualizing the data retrieved at two different periods, such as February 2019 and September 2020. For better understanding, we have conducted an experiment with the top 1000 RePEc²² authors according to the October 2018 rankings. So for each author in that list, identified by the RePEc ID, we are looking for the respective GND ID at WIKIDATA. Initially, there is noted that 994 authors from that list were part of WIKIDATA, and 987 were already mapped with GND ID.

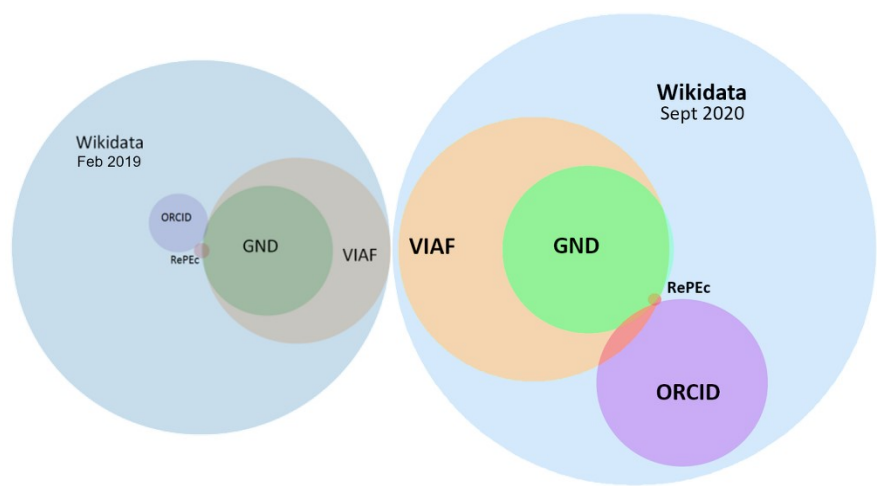


Figure 9.7 Authority (persons) identifiers in WIKIDATA

²² The data are crawled from RePEc (<https://ideas.repec.org/top/top.person.all.html>, accessed 17.10.2018).

9.2.4 The overview of outcomes

The benefits of identifying authors with a respective ID are numerous and in many directions. Firstly, referring to an author with a relevant identifier, besides serving the clustering process within a repository, leads to the generation of an accurate and comprehensible list of publications. Minimizing the doubt as to whether or not a publication belongs to the selected author. This also contributes to the creation of a truthful co-authorship graph.

In addition, the list of discovered and extended author's identifiers, such as VIAF ID, GND ID, WIKIDATA ID, RePEc ID, are used for further data enrichments by harvesting the corresponding sources. The implementation of linked data principles makes it easier to exploit these data by avoiding some complex ETL (extract-transform-load) processes. In such a manner, by consuming DBpedia, GND/Entity Facts and WIKIDATA, the author profile is extended with information such as life data, short abstract, a picture, professions, affiliations, other identifiers, etc. Section 9.2.3 presents some of the listings for this purpose, with the results shown in Figure 9.8.

As a result of these approaches, the generated profile provides to the user the possibility to have an overall view about a certain author, by avoiding multiple navigations to different sites for data collection. Thus, in our selected example the author profile is enriched with data such as other name alternatives, birth year, picture, abstract, affiliations, professions, new publications that are missing in the initial DL, five new co-authors, and linked sources to other libraries. In addition, other profile IDs such as RePEc, SSRN, Twitter, Quora²³, ResearchGate, are displayed.

Figure 9.8 gives a general overview of how such an enrichment may be offered. Moreover, the co-author graph and the word tag cloud with the most frequently used terms and topics are shown, as it is described in figure 9.9 and 9.10, respectively.

²³ <https://www.quora.com>

9. Evaluation of author's disambiguation

ECONSTOR | Authors

PROTOTYPE | Cross-linking author's information from other repositories.

select an author

Sims, Christopher A.

1. Bayesian methods for dynamic multivariate models (1996) [Link Econstor](#)

2. Does monetary policy generate recessions? (1998) [Link Econstor](#)

3. Error bands for impulse responses (1995) [Link Econstor](#)

4. Fiscal Aspects of Central Bank Independence (2001) [Link Econstor](#)

5. MCMC method for Markov mixture simultaneous-equation models: a note (2004) [Link Econstor](#)

6. Methods for inference in large multiple-equation Markov-switching models (2006) [Link Econstor](#)

7. Rational inattention: a research agenda (2005) [Link Econstor](#)

8. Were there regime switches in U.S. monetary policy? (2004) [Link Econstor](#)

9. When does a central bank's balance sheet require fiscal support? (2014) [Link Econstor](#)

1 **Zha, Tao**

A Gibbs simulator for restricted VAR models(2000) [Link Econstor](#)

Assessing simple policy rules: a view from a complete macro model(2000) [Link Econstor](#)

Bayesian methods for dynamic multivariate models(1996) [Link Econstor](#)

Conditional forecasts in dynamic multivariate models(1998) [Link Econstor](#)

Confronting model misspecification in macroeconomics(2012) [Link Econstor](#)

Do credit constraints amplify macroeconomic fluctuations?(2010) [Link Econstor](#)

Error bands for impulse responses(1995) [Link Econstor](#)

Indeterminacy in a forward-looking regime-switching model(2007) [Link Econstor](#)

Land prices and unemployment(2013) [Link Econstor](#)

Land-price dynamics and macroeconomic fluctuations(2011) [Link Econstor](#)

Learning, adaptive expectations, and technology shocks(2008) [Link Econstor](#)

Likelihood-preserving normalization in multiple equation models(2000) [Link Econstor](#)

Liquidity premia, price-rent dynamics, and business cycles(2014) [Link Econstor](#)

MCMC method for Markov mixture simultaneous-equation models: a note(2004) [Link Econstor](#)

Markov-switching structural vector autoregressions: theory and application(2005) [Link Econstor](#)

Methods for inference in large multiple-equation Markov-switching models(2006) [Link Econstor](#)

Minimal state variable solutions to Markov-switching rational expectations models(2010) [Link Econstor](#)

Modest policy interventions(1999) [Link Econstor](#)

Sources of macroeconomic fluctuations: A regime-switching DSGE approach(2010) [Link Econstor](#)

Structural vector autoregressions: Theory of identification and algorithms for inference(2008) [Link Econstor](#)

The conquest of South American inflation(2006) [Link Econstor](#)

The dynamic striated Metropolis-Hastings sampler for high-dimensional models(2014) [Link Econstor](#)

2 **Waggoner, Daniel F.**

A Gibbs simulator for restricted VAR models(2000) [Link Econstor](#)

Asymmetric Expectation Effects of Regime Shifts and the Great Moderation(2007) [Link Econstor](#)

Confronting model misspecification in macroeconomics(2012) [Link Econstor](#)

Density-conditional forecasts in dynamic multivariate models(2010) [Link Econstor](#)

Generalizing the Taylor principle: Comment(2008) [Link Econstor](#)

Indeterminacy in a forward-looking regime-switching model(2007) [Link Econstor](#)

Inference based on SVARs identified with sign and zero restrictions: Theory and applications(2014) [Link Econstor](#)

Likelihood-preserving normalization in multiple equation models(2000) [Link Econstor](#)

Methods for inference in large multiple-equation Markov-switching models(2006) [Link Econstor](#)

Minimal state variable solutions to Markov-switching rational expectations models(2010) [Link Econstor](#)

Normalization in econometrics(2004) [Link Econstor](#)

3 **Del Negro, Marco**

Aggregate unemployment in Krusell and Smith's economy: a note(2005) [Link Econstor](#)

Asymmetric shocks among U.S. states(2000) [Link Econstor](#)

Country versus Region Effects in International Stock Returns(2003) [Link Econstor](#)

Inflation in the great recession and new Keynesian models(2013) [Link Econstor](#)

Monetary policy analysis with potentially misspecified models(2005) [Link Econstor](#)

On the fit and forecasting performance of New Keynesian models(2004) [Link Econstor](#)

Policy predictions if the model doesn't fit(2004) [Link Econstor](#)

The FRBNY DSGE Model(2013) [Link Econstor](#)

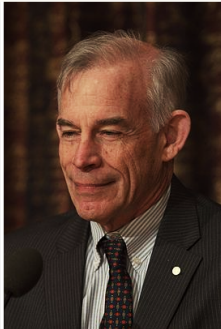
The forward guidance puzzle(2012) [Link Econstor](#)

Time-varying structural vector autoregressions and monetary policy: A corrigendum(2013) [Link Econstor](#)

When does a central bank's balance sheet require fiscal support?(2014) [Link Econstor](#)

Christopher A. Sims

Christopher Albert Sims
Christopher Sims
Chris Sims



1942 Washington, DC

Christopher Albert "Chris" Sims (born October 21, 1942) is an American econometrician and macroeconomist. He is currently the John J. F. Sherrerd '52 University Professor of Economics at Princeton University. Together with Thomas Sargent, he won the Nobel Memorial Prize in Economic Sciences in 2011. The award cited their "empirical research on cause and effect in the macroeconomy".

Profession


Econmist


Affiliations


- Princeton University. Department of Economics

External links

- Gemeinsame Normdatei (GND) im Katalog der Deutschen Nationalbibliothek
- Bibliothèque nationale de France
- Wikipedia (Deutsch)
- Wikipedia (English)
- NACO Authority File
- Virtual International Authority File (VIAF)
- International Standard Name Identifier (ISNI)
- Wikidata

 Google Scholar

 RePEc

 Nobel

Prizes in Economics

2011 - Nobel Prize in Economics

Figure 9.8 An enriched/extended author profile

142

9.2 The outcome after identification



Figure 9.9 Co-authors network for a particular author

The right identification of an author in a DL ensures to us a truthful list of co-authors, associated with that author within the same DL, and a key point for harvesting the missing co-authorship relations from the newly discovered sources. The graphical representation of such relationships reflects a direct view of the authors' collaboration and facilitates users' navigation through them. Figure 9.9 shows an example of such kind of visualization through a word tag cloud, where the font size represents the frequency of co-authorship with a link to the corresponding author profile.

A comprehensive and accurate list of publications is also an important source for further processing and insights. By taking the titles, abstracts and keywords/subjects of publications, we generate a visual view of the most frequent terms and concepts used by the author. Such kind of visual representation provides an instant overview of the covered topics and fields in the author's research output. Figure 9.10 denotes a view of such representation, where the tag cloud represents the most used terms and concepts in the "Oliver D. Hart" publications. Such a view is based on the titles, abstracts and subjects from the current list of his publications indexed in the EconBiz. Any changes to that list, by adding or removing publications, would also affect the following view. In addition to this, the extracted terms can be used by the user for query formulation in an attempt to narrow down the results. Section 6.4 provides more details in this

9. Evaluation of author's disambiguation

concert; even, in that case, the process is based on a single publication, in comparison to all author's publications from the author as it is here.

The set of terms pulled out from authors' publications titles and abstracts, in combination with assigned thesauri subjects, embody an important component to calculate the similarity between authors. Namely, to retrieve the list of authors whose research output intersects with the selected author. As explained in section 6, the assigned descriptors based on a predefined thesaurus, such as the usage of STW thesaurus in our case, are of key importance, considering here the manually labeling process done by the domain experts. The alignments between thesauri concepts play an essential role, especially when information retrieval targets multiple repositories, and especially when it comes to multilanguage or domain-specific environments. However, the provided experiments show to us that the usage of other metadata components, such as title and abstract is necessary to narrow down the results and improve the similarity calculations. Figure 9.11 shows an example of that how the similarity between authors can be exposed to the users. The proposed prototype solution provides also options for some adjustments in order to generate a better result.



Figure 9.10 Word Tag cloud with terms from author's publications

9.2 The outcome after identification

Accordingly, the slider movements imply different compilation of concepts including the corresponding frequencies for the respective author's research output. The slider degree instantaneously determines the level of thesaurus usage in terms of narrowing/broadening the concepts, and the depth of machine learning generated concepts participating in the similarity calculations. Furthermore, the user can choose between the thesaurus and title/abstract concepts inclusion in the calculation. Such that, the following set of concepts: "*contract theory*", "*theory firm*", "*corporate governance*", "*incomplete contract*", "*capital structure*", "*bankruptcy procedure*", "*contracts*" and "*bankruptcy*", gives the results as in the figure 9.11, by considering the example for the author "Oliver D. Hart".

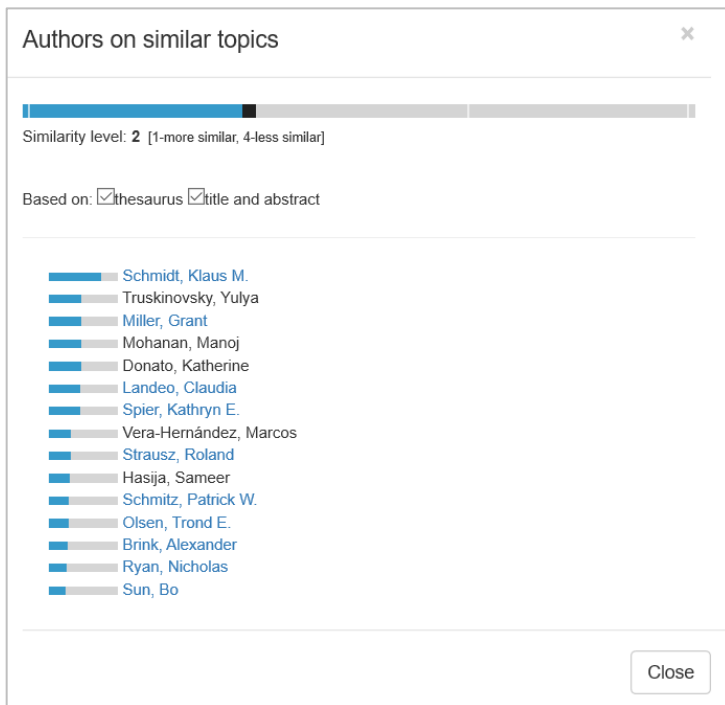


Figure 9.11 Listing authors working on similar topics

9. Evaluation of author's disambiguation

9.3 Summary

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 authors in the initial repository. Besides, except VIAF, WIKIDATA provides a powerful hub of linking authorities. By discovering a persistent identifier about the author, the list identifiers can be expanded with several others, provided by WIKIDATA. Note that our approach 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.

As a result of the applied approach, the generated author profile within a digital library can serve in different directions, by considering that in a single place scholars can get information, such as:

- biographic details, i.e. birth year, affiliations, professions, etc.,
- accurate topic representation based on the research output,
- other publications differ from them within the DL,
- more comprehensive co-authors relations,
- author-related WIKIDATA or DBpedia content,
- recommendation about similar authors or publications,
- social media content.

The outcomes of this approach are successfully integrated at Leibniz Information Centre for Economics (ZBW). Through a very visible platform such as EconBiz, we are providing a wider profile of authors. The running application²⁴ has been evaluated in real environments and positively accepted by the research community and authors that find themselves on it. Therefore, the transfer of knowledge from this approach shows that the adopted techniques work really well in a real environment.

²⁴ <https://authors.econbiz.de>

Related Work

*“Any fact becomes important
when it's connected to another”*

Umberto Eco

A very good part of the related work has been mentioned in various parts of the relevant sections. Here is a summary of related work that is linked to the two main approaches followed in this work.

10.1 Recommender systems across LOD

10.1.1 Linked Open Data in Recommender Systems

The implementation of semantic technologies and the approach of interlinking resources known as Linked Data has given a new vision to the interoperability among information [BHLO01]. Since the conceptualization of Linked Data principles in 2006, as a set of best practices for publishing and interlinking structured data on the Web, the intention of them has been increased rapidly [AHBB16]. The RDF data model appears to be a widely accepted model for data integration, knowledge representation, and interconnections. Due to this, Digital Libraries often prefer to publish their indexes or even entire catalogs as RDF serializations. This intention does not rely only on publishing; applying and consuming Linked Data principles in real applications is now a common practice. Among several examples, a remarkable one is Europeana; an aggregator and single access point to millions of books, paintings, films and museum objects [DGH10]. Alignments of concepts i.e. SKOS mappings among repositories/thesauri

10. Related Work

can play a crucial role in the process of interoperability and interdisciplinary. The ARIADNE project highlights the importance of vocabulary linked data for the integration of archaeological records [BiTu16]. Several other projects put the focus on querying and retrieving information from LOD based on these alignments [FCLV11, JJHY12].

The usage of DBpedia in the context of recommender systems has been seen with high relevance, as it represents the nucleus of the LOD cloud. There are present a large number of researches in this direction, where the implementation of DBpedia content is used for semantic similarity measurement in graph-based recommendation [MBLG17, MeDa15, Pass10a]. The implementation of LOD in recommender systems is present in several domains, such as in the domain of music [Pass10b], scientific publications [HaLT14], book recommendation [PeVo13], movies [VFRT16], etc. In most of these cases, the DBpedia properties are used for semantic similarity calculation. In addition, the usage of Freebase is debried as a possibility for enriching artists with other related content, also [BaSc12]. The explorations in this field are in continuous progress. Hence, in very recent research is introduced a generic framework based on linked open data, that is algorithm-independent and domain-independent [MNLG18]. It can generate a natural language explanation for every kind of recommendation algorithm, as investigated in three different domains, like movies, books and music.

Retrieving information relying on the linked data knows to generate a very high recall [CuLL15]. Usually, the result is dominated by the information that can be so different from what the user is interested in, i.e. not relevant to the user, or any relevant information cannot be displayed. Providing the user with the desired information, several parameters must be considered, such as the previously selected item or any other kind of preference [NMOR12]. Such that, it is inevitable to explore the application of recommender systems in scholarly communication, particularly in digital libraries [HCOC02, MoRo00, SmCa05]. The common implementation of recommending systems in DLs is mainly a practice used within the same repository. Therefore, recommending and interlinking publications by crosslinking relevant information from several repositories remains a challenge [DSEQ13, Hora10, Pass10a]. The systems for retrieving and recommending scientific publications are generally grounded on

content analysis, user profiles and collaborative filtering with the incontestable role of social data [BOHG13, LoGS11, PKCK12, SuKa10]. Therefore, the application of data mining approaches is necessary.

10.1.2 Vector Space Model and Word Embedding

The implementation of the Vector Space Model is one of the most propagated approaches in information retrieval, collaborative filtering and recommender systems [LoGS11, Must10, PaBi07]. According to Turney and Pantel [TuPa10], the calculation of word frequencies and weighting elements can guide machines to understand the meaning of human language. TF-IDF is one the most popular weighting scheme, with around 70% of text based recommender-systems [BGLB16].

Consequently, the Vector Space Models (VSMs) of semantics are beginning to address these limits mainly through TF-IDF and Cosine Similarity. The truncated Singular Value Decomposition (SVD) based on Deerwester et al. [DDFL90] can be applied to document similarity, known as Latent Semantic Indexing (LSI), or Latent Semantic Analysis (LSA) in case of word similarity. According to Zelikman, the introduced Contextual Salience (CoSal) can represent a replacement for TF-IDF and an intuitive measure of contextual word importance [Zeli18].

As emphasized in the previous chapters, the lexical features, like string matching and frequency of words in a text, do not capture semantic similarity at a satisfactory level [BaDK14, KeRi15]. In fact, this approach gives a satisfactory precision in terms of harvesting and ranking publications. However, on the other hand, many relevant documents remain very low ranked, only because there is not a match between the proposed query and their metadata.

Current trends for determining word similarities, i.e., semantic similarities among texts, rely on vector representations of words by using neural networks, known as *word embedding* or word representations [BaDK14, BSSM06, TuRB10, CoWe08, KeRi15, KSKW15, LeCo15, LeGD15, MCCD13, MnHi09, PeSM14]. Hence, word embedding has found an extended application in areas such as recommender systems [BaKo16, MSGL16], information retrieval [AmMG16, GRMJ15], sentiment analyses

10. Related Work

[TWYZ14], document clustering, bilingual machine translation [ZSCM13], etc. As shown in section 4.2.2, Word2Vec is a very popular word embedding approach, which learns a vector representation for each word using the neural network language model [MCCD13]. In addition, there are several other customizations and implementations. Such as an example of learning text representations is FastText [BGJM17, JGBM16]. Recently Athiwaratkun et al. introduced the Probabilistic FastText, as a new model for word embedding that can capture multiple word senses, sub-word structure, and uncertainty information [AtWA18]. Besides, there are several approaches that are mixing the existing popular algorithms with word embedding approaches. The mixture of Dirichlet Topic Models and Word Embedding, i.e. word2vec, into Ida2vec represent a case [Mood16].

However, in several publications, the efficiency, performance, as well as shortcomings and limits of the word embedding approach are tackled. Therefore, the meaning of the similarity values and the lack of an intuitive threshold for a given embedding mode, are mentioned by Elekes et al. [EESB18]. Several other publications also emphasizing that the values of parameters and the corpus selection during the training phase are influencing the relatedness and similarity of the generated word embedding vectors [FTRD16, HSMN12, LeGD15, LLHZ16, SLMJ15, YSMB16, ZaCr16]. Mikolov et al. in a recent publication present promising approaches for training high-quality word vector representations by using a combination of known tricks that are however rarely used together [MGBP17].

10.2 Authors' disambiguation and identification

Determining if a particular research output belongs to a specific author, remains a permanent challenge in the world of libraries. Author name ambiguity is a foremost problem which at any moment raises doubts as to whether or not a work belongs to a particular author, having in mind the cases where distinct authors have the same name or the same author may be represented with different names. Consequently, the need for unique author representation may not have been posed as a major problem at a

10.2 Authors' disambiguation and identification

time when data are stored in a singular repository, where it was not possible and necessary to link or share the information with other repositories. However, at the moment when it is intended to interlink information, accurate identification of persons, but not only, is a crucial need. Hence, in order to be able to perform an author-targeted query with satisfactory precision and recall, retrieving all and only those publications by a particular author, the authorship records for this author need to be disambiguated [KiDi16, LIKC14, MüRR17, Salo09].

10.2.1 Authors disambiguation

The main challenge in the disambiguation process is the identification of whether two authors in the same or different DLs have the same identity or not. Generally, two main steps are applied for this purpose, measuring the similarity and clustering similar records. However, various studies are proposing different strategies and algorithmic approaches. The most explored strategies consider the string processing approach which measures the similarity of authors' names [BMCR03, ToSm09]. The comparisons are one-to-many and many-to-many, by applying iterative [BhGe04] and incremental methods [SGLF17]. The explored disambiguation approaches are generally divided into supervised with heuristic similarity functions, unsupervised and hybrid [FeGL12, TFWZ12].

In almost all these strategies, the author's disambiguation process is primarily based on relationships among co-authors and similarity of publications, by discovering other relationships in other DLs [FWPZ11, KNLJ09]. Several approaches are relying on the citation network for clustering authors with different algorithmic methods [GiZH05, MaYa03, SmTo09]. The most applied techniques for disambiguation and clustering are based on k-way spectral clustering [GiZH05], latent topic-based approaches (as an extension of Bayesian text model, such as the extension of Latent Dirichlet Allocation) [SHCL07, ShLM09], Support Vector Machines (SVMs) and the vector space representation [HGZL04], by including algorithms such as the Cosine Similarity (CS) with TF-IDF, Jaccard Similarity, Jaro Winkler, and Levenshtein algorithms. There are cases when the gathered information from citations is processed and used

10. Related Work

directly in Web search engines to find relevant information about authors [PRZL09].

In addition, user feedback is highlighted as a valuable opportunity to facilitate and enhance the accuracy of the disambiguation process [FeMG12, FWPZ11]. Therefore, in several publications is emphasized that the feedback in combination with the hybrid supervised process is applied for assigning references to authors [FeMG12, GTCG13, TFWZ12]. Especially the process can benefit from the voluntary participation of active authors [LIKC14].

10.2.2 Linking Authors

The introduction of linked open data offers new opportunities for discovering and interlinking authors and their research outputs [NeTo12]. Nowadays there are several efforts for generating authority profiles by aggregating and uniquely identifying resources and researchers, such as ORCID, VIAF, ISNI, VIVO, Google Scholar, Scopus, Mendeley, Academia.edu, Microsoft Academia, ResearcherID, OpenID, etc. According to [NeTo12] and several other studies, authority files are valuable sources that offer a backbone for the Semantic Web. Section 2.3 highlights and gives more details about these services by underlining VIAF and WIKIDATA.

In addition, a number of ongoing projects are focused on persistent identifiers (PIDs). In this context, we distinguish FREYA²⁵ project, which aims to extend the PIDs infrastructure, and facilitate the open research in EU countries and globally. Hence, the interconnection of identifiers affects the improvement of data interoperability i.e. discovery, navigation and retrieval of research resources [WiFe18].

²⁵ <https://www.project-freya.eu>, accessed 23.01.2019

Part IV

Conclusion and Future Work

Conclusion

*“there is no unique picture
of reality”*

STEPHEN HAWKING

DLs are limited to certain institutional “barriers” – collections, metadata, services, etc. – that researchers can rely on. As a consequence, starting from a specific DL it is impossible or at least very difficult to cross the boundaries by spreading one’s search to other resources. These features are important concerns when a scholar or an experienced researcher wants to get insights into a new research field, the author’s new publications, or their relevance.

As expressed in the motivation part, the significance of DL is key to effective scholarly communication. Their usage brings considerable benefits to the research community, enabling global and simplified access to scientific resources (publications, research data, etc.).

Thus, by crosslinking data from different places, a resource would be enriched with additional library or non-library resources. This results in a significant enhancement of scholarly communication, i.e. a more efficient information retrieval process. The idea is to perform a single query in a single place (e.g. their favorite DL) and offer scholars information from different repositories. Ultimately, a selected publication in a DL will be enriched with a list of recommended publications from other DLs, additional information about authors, conferences, etc.

For this purpose, two main approaches are followed, as described in section 5.3. The first one uses available publication-centered metadata (see section 5.4), while the second one considers available author-centered metadata (see section 5.5).

11. Conclusion

The concept of Linked Open Data presents a broader vision for information exchange and the mitigation of barriers between repositories. Therefore, the aligned terms between repositories (thesauri) are considered an important linking point for publications from different repositories (see figure 6.2). The initial evaluations are performed based on these alignments, and evaluated results show that:

- retrieving semantically similar publications from other repositories based on terms alignments is wide-ranging (e.g., very different results based on mapped terms). Harvesting publications from a targeted repository using aligned terms results in an empty set since the variability of terms used in different repositories for describing a publication. Section 6.1 gives in-depth analysis and several scenarios considering this issue.

The presence of the thesauri in the source and/or the target repository gives another dimension to the interlinking process. Except for the terms that are aligned, the set is extended with narrowed, broadened and related concepts through the Simple Knowledge Organization System Reference - SKOS modeling scheme (see table 6.1).

- However, the impact of such enforcement does not directly affect the retrieving process by narrowing the list of results or by improving the semantic similarities of publications. Such an extension is reflected in the enrichment of the terms with supplementary, semantically-related concepts, necessary for getting a subset of publications from the targeted repository. Figure 6.4 provides an overview of this.
- Another aspect of this operation applies to further steps, as a component for determining the weight of the terms in the metadata set. To this end, the role of the aligned terms between repositories is of particular importance for the crosslinking phase between repositories as well as the extension of concepts extracted from the publication's metadata.

In addition, to narrow down the results and improve the semantic relatedness between the triggered and retrieved publications, the applications of IR methods are considered. To this end, the Vector Space

Model (VSM) and Word Embedding (WE) approaches are deployed and analyzed comparatively. Section 6.2 and 6.3 explain the implementation and experimental setups of these approaches. Consequently, for operating with these approaches, except the thesauri descriptors, the other metadata components are introduced. Concerning the open access policy to most of the documents, our focus is concentrated on metadata such as the title, abstract, authors, co-authors and keywords, without including the full text.

The evaluations are done in crossdomain repositories (i.e., economics vs. agricultural), and the proposed approaches are analyzed comparatively with different sets of metadata. In order to discover more insights, the analyses are performed in unlabeled datasets, hence on several occasions, the human assessment is applied. Section 8.1 provides all the details considering the evaluations followed by this strategy. To this end:

- regarding the number of relevant documents in the top ten, the results show a slight superiority of word embedding through Word2Vec implementation. However, TF-IDF and CS achieve a better ranking of the documents in that list. Therefore, the traditional count-based and text-matching approach of VSM is achieving almost similar results such as the word embedding approach, when a large set of metadata is considered (i.e. title, abstract, keywords, and descriptors).
- in the case when a reduced amount of metadata is measured, for example only the publication title, the WE approach outperforms the previous approach.
- both approaches perform differently considering the top 10 relevant publications. That list of publications based on VSM is 40.5% different from the list generated through the WE approach.
- TF-IDF and Cosine Similarity rely on the exact match among the compared publications. The inability to determine word relatedness seems to be the main weakness, which affects this approach with regards to ranking results if they do not have common words. Even the application of external thesauri and vocabularies does not provide any visible improvement if a human is not involved. On the other hand, the WE reaches impressive relatedness among terms, which impact in having as top-ranked publications with few or

11. Conclusion

almost null exact matched words. However, this approach is affected and is sensitive to the chosen datasets and the values of the predefined parameters for training the model. These elements affect the quality of the model and its performance in real cases.

Based on the experimental setup and evaluations, choosing the right metadata is a crucial element in the process of crosslinking publications. The publication metadata selection from where the search is initiated is considered as bait for successful “fishing,” i.e., retrieving more relevant information from other repositories. Through the combination of metadata elements (title, abstract, and keywords) with the thesauri descriptors, the overall terms are weighted and ordered according to their importance. However, not all terms created in this way are necessary for subsequent calculations. In that set, we may have more general terms that can mislead the results – i.e., retrieve publications that are semantically distant from the initial publication. Thus, an important role is to prioritize their selection according to their weight and meaning. That is also important, when only a small set of terms are available, for example, determining the semantic similarity between publications based solely on the titles.

- Among the methods we experimented with, the TF-IDF identifies the most important terms in cases when the right combination among the metadata is selected. Hence, the presence of abstract, keywords, thesaurus descriptions and doubling the weight of the title metadata element, gives the best combination. When considering the crosslinking process (retrieving semantically similar publications) from repositories of a different domain, the definition of term weights is more effective if it is made using global terms frequencies than those generated by the initial corpus.
- The possibility for users to perform manual adjustments of these metadata components (title, abstract, keywords, and descriptors) - by viewing the impact of the terms during a search in real-time increases the quality of the retrieved publications.

Analyzing the generated results as well as the fact that the above approaches act differently from one another, a possible linkage of these

approaches to a single interface shows significant improvements. Such an implementation is presented in more detail in section 6.4.

- the application of visual search interfaces is proposed as a way to simplify and provide a more intuitive retrieval of similar publications based on a preselected publication. This enables the scholar to perform more detailed research with a reduced mental workload, in comparison to traditional keyword-based search. The proposed approach, in an innate and conceptual manner, makes possible the application of suggested terms from other external resources. Accordingly, the set of terms can be extended with terms or concepts from an external language thesaurus, any SKOS modeling scheme, and at the same time, the deployment of terms through machine learning techniques applied innately. This allows the scholar at any time to manage the features and instantly see the change reflected on the results.

The second approach is referred to the process of enriching the profile of an author inside a digital library by harvesting available information from other repositories. To this goal, the prior disambiguation of authors plays a decisive role. This issue is mainly tackled in chapter 7 through an algorithmic approach, including VIAF usage, while the outcomes and evaluations are described in chapter 9, specifically in section 9.2.

- Starting with a set of metadata that usually describes an author, such as the name, publications and a list of co-authors, an author can be identified with high precision in VIAF. The F1 measures for this can reach a score of 0.975. Therefore, the application of authority files such as the virtual international authority file – VIAF, integrated authority file (GND), and even WIKIDATA, significantly improves the disambiguation process.
- Such identification helps us assign a globally known identifier to the author by facilitating the harvesting process in the following steps. Moreover, WIKIDATA and VIAF are also considered as hubs that support the further enrichment of results.

The transfer of knowledge from this approach is used and assessed in a real-world environment through the EconBiz platform. The application

11. Conclusion

provides a comprehensive profile of authors by interlinking and integrating data from various sources. Such that, in a single interface the scholar can find a broad view of information in regard to a selected author, e.g. the most prominent fields of research the author is engaged in. The evaluations are done with scholars and authors themselves, find such an approach very useful from different points of view. The scholars value the volume, relevance, and diversity of information in one place. Moreover, the ability to find new insights quickly and intuitively is emphasized. The authors appreciate that they do not need to compile the information themselves as is often required by other author profile services. This approach also provides support to DLs for a better quality check and data curation. By visualizing the aggregated and clustered information, it enables the identification of some issues such as duplicates, false authorship, etc.

Future Work

*“it matters if you
just don’t give up”*

STEPHEN HAWKING

The thesis implements and evaluates several approaches, methods, and algorithms for linking and enriching publications and authors in one DL with information from other DLs. Such kind of linking consists of semantically relevant publications related to a particular publication, co-author network list, author-related information for creating or enriching her profile, etc. Numbers of approaches like semantic web, linked open data consumptions, i.e. thesauri concept alignments, data mining and machine learning techniques were explored to achieve the objectives, however, there are still several other approaches that may be explored in the future.

One aspect that can be addressed as future work is the evaluation of the output generated as a result of the first strategy, described in chapter 8. Currently, we offer an ultimate prototype from where the user can assess the relevance of the recommended publications based on each approach. However, as explained in chapter 8, this method is time-consuming and challenging in concern to subjects’ engagement, i.e. users for performing evaluations. Therefore, the implicit feedback generated from click-through [JGPH05], would increase the number of evaluations and also would offer facilitation in the evaluation process, especially if the application is based on a real environment [KnWi15]. For example, for a certain publication that the user has selected from the recommendation list, which approaches contributed to its ranking, how long the user stayed on it, etc. A setup that will act in the background for observing user behaviors i.e. decisions on item selections also presents an opportunity to extend the range of

12. Future Work

experiments and improvements in the process of similarity measurement. This will be done through the combination of several metadata elements such as title, keywords, thesauri subjects, abstract with the already explored approaches and by introducing new methods mainly from the field of machine learning. Another point where we will focus in the future is on consuming the concept alignments between repositories or thesauri, i.e. expanding the scope of operations in other domains such as medicine, computer sciences, social sciences, etc. Hence, the crossdomain linkage and information retrieval will get a more comprehensive view. Moreover, in addition to the similarity and accuracy to apply and evaluate the diversity and novelty at DL resources.

Considering the second strategy, of disambiguating, linking and generating profiles about authors, Wikidata will continue to be in the center of our attention. Actually, the applied approaches for disambiguating and assigning a PID to authors, by using VIAF, GND and WIKIDATA are resulting in very satisfying outcomes. However, the permanent growth of WIKIDATA, where the presence of such identifiers is increasing constantly (see table 2.1, fig 9.7), position it to a crucial hub for linking authorities (fig. 9.6). Current practices, where a number of services are being fed with data from WIKIDATA, such as Scholia, Entity Fact, indicate its crucial role in information aggregation. Undoubtedly, the contribution of the community in data gathering and updating emphasizes further improve these attributes. All this gives strong support and provides an important momentum to the open science movement.

Part V

Appendix

Definitions and the Algorithmic Approach

A.1 The usage of some definitions in our context

Resource – Refers to an item of a library catalog. In our context, it represents an intellectual output i.e., publication, of one or several authors, stored in a particular repository or DL. A specific resource is represented and described through a bibliographic record. Authority records are used to achieve better consistency of bibliographic records and to provide a linking structure between them.

Repository – Provides storing and publishing of resources, i.e. intellectual output, from institutions, universities, organizations, etc. It is very common for them to be categorized in specific domains, such as economics, medicine, computer sciences. Besides the resource metadata, the repository usually stores the full-text, for the cases where the content adheres to the Open Access principles. Moreover, there are also several cases where a part or entire repository content is provided as a dump file, a SPARQL endpoint, APIs, or through a search interface. Such an example is the EconStor repository that offers full-texts (working papers, journal articles, conference proceedings, etc.) from the business administration and economics domain. Its content can be accessed from the search interface but also part of it is available as a dump file and from the SPARQL endpoint. However, not all repositories always offer these access options; the lack of a search interface is noted in several cases.

A. Appendix

Dataset – by definition, it presents a collection of data that usually are logically integrated. We are referring to datasets also for a part or entire repository content. In general, a repository may also be seen as a collection of several datasets.

Digital Library – in most cases a DL represents a kind of gateway for different resources that do not necessarily belong to a repository. However, a DL may also represent the interface for a single repository. Furthermore, in a DL, a resource can be presented by only its metadata such as title, authors, keywords/subjects and abstract, without providing the full-text. In comparison to a repository, a DL offers a search interface, including faceting possibilities.

Since all of our proposed approaches are based and operate on the descriptive metadata layer e.g. title, keywords, authors and abstract, the use of both definitions, namely repository or DL, may be applicable. For this reason, we are referring to both concepts in the text. In fact, most of our experiments and evaluations are done on repository subsets, such as EconStor and AGRIS, which for practical purposes are loaded in our local environment. However, we are referring to a DL as a broader concept, since the proposed approaches and findings are not limited just to the repository level. They are applicable in a DL or repository when the metadata set such as title, authors, keywords and the abstract are provided.

Initial repository / DL – refers to the repository or DL from where the scholar initiates the search. It also refers to the repository/DL whose content is enriched and linked with information from other places (repositories/DLs). For example, if we consider EconStor as an initial repository, then its resources are linked with semantically related resources from other repositories, authors are enriched with additional information by generating a distinct profile, etc.

Target repository / DL – The repository or DL we target as a place to retrieve the required information from, such as semantically similar publications, additional information about authors, etc.

User / End-user / Scholar – a person that interacts with the repository/DL search interface.

Term / Concept – a term, i.e. a KOS term, characterizes a particular concept. Therefore, in the cases when the concept is atomic in fact it represents a term, hence in this work are used interchangeably. However, the relationship between the term and the concept is more complex. Typically in natural languages, different terms are used to describe a particular concept, known also as labels, and not in all the cases the same term is used for the same concept. Hence, one of the primary roles of thesauruses is also to organize concepts through different relationships among terms, such as the hierarchical approach over broader and narrower terms.

Descriptor / Subject heading – represents the preferred term (label) for describing a concept, while the non-descriptor terms are represented through alternative labels.

A. Appendix

A.2 The algorithmic approach to disambiguate authors through VIAF clusters

Input data from a digital library: \mathbf{a} , P_a , A_a and \bar{P} .

Input data from a VIAF cluster: A_{cj} , P_{cj} , \hat{A}_{cj} and \check{P}_{cj} .

Output: determine the match between the authors \mathbf{a} with a VIAF cluster c_j .

Similarity measurement for point (i) with the equation (7.1)

```
1:   if  $A_{cj} \neq \emptyset$  then
2:     for each  $a_i^{cj} \in A_{cj}$  do
3:       if cosine_distance( $a, a_i^{cj}$ ) = 1 then
4:          $w_{ac} = 0.5$ ; (the max value of  $w_{ac}$  can be 1)
5:       end if
6:     end for
7:   end if
```

Similarity measurement for point (ii) with the equation (7.2)

```
8:   if  $P_a, P_{cj} \neq \emptyset$  then
9:     for each  $p^a \in P_a$  do
10:      for each  $p^{cj} \in P_{cj}$  do
11:        if ( $p^a, p^{cj}$  have more than 3 terms in title) then
12:          if cosine_distance( $p^a, p^{cj}$ )  $\geq 0.9$  then
13:             $w_{pc} = +2$ ;
14:          else if cosine_distance( $p^a, p^{cj}$ )  $\geq 0.6$  then
15:             $w_{pc} = +0.5$ ;
16:          end if
17:        end if
18:      end for
19:    end for
20:  end if
```

Similarity measurement for point (iii) with the equation (7.3)

```
21:  if  $A_a, \hat{A}_{cj} \neq \emptyset$  then
22:    for each  $a^a \in A_a$  do
23:      for each  $\hat{a}^{cj} \in \hat{A}_{cj}$  do
24:        if cosine_distance( $a^a, \hat{a}^{cj}$ ) = 1 then
25:           $w_{\hat{a}_c} = +2$ ;
26:        else if jarowinkler_distance( $a^a, \hat{a}^{cj}$ )  $\geq 0.9$  then
27:           $w_{\hat{a}_c} = +0.5$ ;
28:        end if
29:      end for
30:    end for
31:  end if
```

Similarity measurement for point (iv) with equation (7.4)

- 32: Calculations in this step are absolutely the same as in point (ii).
 Instead of the set P_{c_j} is used \check{P}_{c_j} with its elements. An initial step,
 check if $(p^a \neq p^{c_j})$ is true, perform the calculation. This condition
 avoids the same publication to be measured more than ones. The
 weight for the step is denoted with w_{ps} .

Determine the matching result

- 33: **if** $(w_{pc} + w_{ac})$ or $(w_{\hat{a}_c} + w_{ac})$ or $(w_{\check{p}_c} + w_{ac}) \geq 2.5$ **then**
 34: the cluster c_j is assigned as "correct" for the author a
 35: store the VIAF ID in our the database for the author a , as "correct"
 36: **elseif** $(w_{pc} + w_{ac})$ or $(w_{\hat{a}_c} + w_{ac})$ or $(w_{\check{p}_c} + w_{ac}) \geq 1.5$ **then**
 37: the cluster c_j is assigned as "maybe" for the author a
 38: store the VIAF ID in our the database for the author a , as "maybe"
 39: **else**
 40: the cluster c_j is assigned as "incorrect" for the author a
 41: store "NA" instead of VIAF ID in our the database for the author a
 42: **end if**

Bibliography

- [AdTu05] Adomavicius Gediminas, and Tuzhilin Alexander. "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions". In: *IEEE Transactions on Knowledge and Data Engineering* vol. 17 (2005), Nr. 6, pp. 734-749.
- [AgFS16] Agosti Maristella, Ferro Nicola, and Silvello Gianmaria. "Digital library interoperability at high level of abstraction". In: *Future Generation Computer Systems* vol. 55 (2016), pp. 129-146.
- [AgFS18] Agosti Maristella, Ferro Nicola, and Silvello Gianmaria. "Digital Libraries: From Digital Resources to Challenges in Scientific Data Sharing and Re-Use". In: *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years* : Springer, 2018, pp. 27-41.
- [AGHH12] Antoniou Grigoris, Groth Paul, Harmelen Frank van van, and Hoekstra Rinke. "A Semantic Web Primer", *The MIT Press*, 2012. — ISBN 0262018284, 9780262018289.
- [AHBB16] Auer Sören, Heath Tom, Bizer Christian, and Berners-Lee Tim. "Linked data on the web (LDOW2016)". In: *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*. New York, NY, USA: ACM Press, 2016, pp. 1039-1040. — ISBN 9781450341448.
- [AJCS15] Anibaldi Stefano, Jaques Yves, Celli Fabrizio, Stellato Armando, and Keizer Johannes. "Migrating bibliographic datasets to the Semantic Web: The AGRIS case". In: *Semantic Web* vol. 6, IOS Press (2015), Nr. 2, pp. 113-120.
- [AISR12] Alemu Getaneh, Stevens Brett, and Ross Penny. "Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries: A social constructivist

Bibliography

- approach". In: *New Library World* vol. 113 (2012), Nr. 1/2, pp. 38–54.
- [AmMG16] Amer Nawal Ould, Mulhem Philippe, and G ry Mathias. "Toward word embedding for personalized information retrieval". In: *Neu-IR: The SIGIR 2016 Workshop on Neural Information Retrieval*, 2016.
- [ASWF14] Akbar Monika, Shaffer Clifford A., Weiguo Fan, and Fox Edward A. "Recommendation based on Deduced Social Networks in an educational digital library". In: *IEEE/ACM Joint Conference on Digital Libraries* : IEEE, 2014, pp. 29–38. – ISBN 978-1-4799-5569-5.
- [AtWA18] Athiwaratkun Ben, Wilson Andrew Gordon, and Anandkumar Anima. "Probabilistic FastText for Multi-Sense Word Embeddings". In: *arXiv eprint arXiv:1806.02901* (2018).
- [BaDK14] Baroni Marco, Dinu Georgiana, and Kruszewski Germ n. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". In: *ACL (1)*, 2014, pp. 238–247.
- [BaKo16] Barkan Oren, and Koenigstein Noam. "Item2vec: neural item embedding for collaborative filtering". In: *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*, 2016, pp. 1–6.
- [BaSc12] Baumann Stephan, and Schirru Rafael. "Using linked open data for novel artist recommendations". In: *In 13th Internal Society for Music Information Retrival Conference*. Porto, Portugal, 2012.
- [BCBD16] De Boom Cedric, Van Canneyt Steven, Bohez Steven, Demeester Thomas, and Dhoedt Bart. "Learning Semantic Similarity for Very Short Texts". In: *Proceedings of the 15th IEEE International Conference on Data Mining Workshop, ICDMW '15*, 2016. – ISBN 9781467384926.
- [BeMc04] Beckett Dave, and McBride Brian. "RDF/XML syntax

- specification (revised)". In: *W3C recommendation* vol. 10 (2004), Nr. 2.3.
- [Bern00] Berners-Lee Tim. "Semantic Web on XML", *XML 2000 - Slide list*. URL <https://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html> - retrieved 2019-07-17.
- [Bess02] Besser Howard. "The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries". In: *First Monday* vol. 7, Valauskas, Edward J. (2002), Nr. 6.
- [BGJM17] Bojanowski Piotr, Grave Edouard, Joulin Armand, and Mikolov Tomas. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* vol. 5 (2017), pp. 135–146.
- [BGLB16] Beel Joeran, Gipp Bela, Langer Stefan, and Breitingner Corinna. "Research-paper recommender systems: a literature survey". In: *International Journal on Digital Libraries* vol. 17, Springer (2016), Nr. 4, pp. 305–338.
- [BhGe04] Bhattacharya Indrajit, and Getoor Lise. "Iterative record linkage for cleaning and integration". In: *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2004, pp. 11–18.
- [BHIB08] Bizer Christian, Heath Tom, Idehen Kingsley, and Berners-Lee Tim. "Linked data on the web (LDOW2008)". In: *Proceeding of the 17th international conference on World Wide Web - WWW '08*. New York, NY, USA, ACM Press (2008), p. 1265 – ISBN 9781605580852.
- [BHLO01] Berners-Lee Tim, Hendler James, Lassila Ora, and others. "The semantic web". In: *Scientific american* vol. 284, New York, NY, USA: (2001), Nr. 5, pp. 28–37.
- [BiHB09] Bizer Christian, Heath Tom, and Berners-Lee Tim. "Linked data-the story so far". In: *Semantic services, interoperability and web applications: emerging concepts* (2009), pp. 205–227.

Bibliography

- [BiMo03] Bilenko Mikhail, and Mooney Raymond J. "Adaptive duplicate detection using learnable string similarity measures". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*. New York, NY, USA : ACM Press, 2003, p. 39. — ISBN 1581137370.
- [BiTu16] Binding Ceri, and Tudhope Douglas. "Improving interoperability using vocabulary linked data". In: *International Journal on Digital Libraries* vol. 17, Springer Berlin Heidelberg (2016), Nr. 1, pp. 5–21.
- [BINJ03] Blei David M, Ng Andrew Y, and Jordan Michael I. "Latent dirichlet allocation". In: *Journal of Machine Learning Research* vol. 3 (2003), Nr. Jan, pp. 993–1022.
- [BMCR03] Bilenko Mikhail, Mooney Raymond, Cohen William, Ravikumar Pradeep, and Fienberg Stephen. "Adaptive name matching in information integration". In: *IEEE Intelligent Systems* vol. 18, IEEE (2003), Nr. 5, pp. 16–23.
- [BoFu02] Borgman Christine L, and Furner Jonathan. "Scholarly communication and bibliometrics". In: *Annual Review of Information Science and Technology* vol. 36 (2002), Nr. 1, pp. 2–72.
- [BOHG13] Bobadilla Jesús, Ortega Fernando, Hernando Antonio, and Gutiérrez Abraham. "Recommender systems survey". In: *Knowledge-Based Systems* vol. 46 (2013), pp. 109–132.
- [Borg02] Borgman Christine L. "Challenges in Building Digital Libraries for the 21st Century". In: *International Conference on Asian Digital Libraries* : Springer Berlin Heidelberg, 2002, pp. 1–13.
- [Borg10] Borgman Christine L. "Scholarship in the Digital Age: Information, Infrastructure, and the Internet", MIT Press, 2010. — ISBN 0262514907, 978026251490.
- [Borg90] Borgman Christine L. "Scholarly Communication and Bibliometrics". Newbury Park, CA : Sage Publications, Inc, 1990. — ISBN 0803938799.

- [Borg99] Borgman Christine L. "What are digital libraries? Competing visions". In: *Inf. Process. Manage.* vol. 35 (1999), Nr. 3, pp. 227–243.
- [Brat07] Bratt Steve. "Semantic Web, and Other Technologies to Watch". URL <https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb,W3C>. - retrieved 2019-07-16.
- [BrFr06] Brants Thorsten, and Franz Alex. "Web 1T 5-gram corpus version 1.1". In: *Google Inc* (2006).
- [BrGM14] Brickley Dan, Guha Ramanathan V, and McBride Brian. "RDF Schema 1.1". In: *W3C recommendation* vol. 25 (2014), pp. 2004–2014.
- [BrMA16] Brokos Georgios-Ioannis, Malakasiotis Prodromos, and Androutopoulos Ion. "Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering". In: *arXiv preprint arXiv:1608.03905* (2016).
- [BSSM06] Bengio Yoshua, Schwenk Holger, Senécal Jean-Sébastien, Morin Frédéric, and Gauvain Jean-Luc. "Neural Probabilistic Language Models". In: *Innovations in Machine Learning: Theory and Applications* vol. 194. Berlin/Heidelberg, Springer-Verlag (2006), pp. 137–186.
- [CaKe11] Caracciolo Caterina, and Keizer Johannes. "Thesaurus Alignment for Linked Data Publishing". In: *Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011 Thesaurus* (2011), pp. 37–46.
- [CMWS15] Celli Fabrizio, Malapela Thembani, Wegner Karna, Subirats Imma, Kokoliou Elena, and Keizer Johannes. "AGRIS: providing access to agricultural research data exploiting open data on the web". In: *F1000Research* vol. 4 (2015).
- [CoWe08] Collobert Ronan, and Weston Jason. A unified architecture for natural language processing. In: *Proceedings of the 25th international conference on Machine learning - ICML '08*. New

Bibliography

- York, NY, USA: ACM Press, 2008, pp. 160–167. — ISBN 9781605582054.
- [Crow02] Crow Raym. "The Case for Institutional Repositories: A SPARC Position Paper". In: *Scholarly Publishing* vol. 223, The Association of Research Libraries (ARL) (2002), pp. 1–37.
- [CSRM12] Caracciolo Caterina, Stellato Armando, Rajbahndari Sachit, Morshed Ahsan, Johannsen Gudrun, Jaques Yves, and Keizer Johannes. "Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case". In: *International Journal of Metadata, Semantics and Ontologies* vol. 7, Inderscience Publishers (2012), Nr. 1, pp. 65–75.
- [CuLL15] Cuzzocrea Alfredo, Lee Wookey, and Leung Carson K. "High-Recall Information Retrieval from Linked Big Data". In: *IEEE 39th Annual Computer Software and Applications Conference* : IEEE, 2015, pp. 712–717. — ISBN 978-1-4673-6564-2.
- [CyWL14] Cyganiak Richard, Wood David, and Lanthaler Markus. "RDF 1.1 concepts and abstract syntax", URL <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225>, 2014. — W3C Recommendation 25 Feb. 2014.
- [DDFL90] Deerwester Scott, Dumais Susan T, Furnas George W, Landauer Thomas K, and Harshman Richard. "Indexing by latent semantic analysis". In: *Journal of the American society for information science* vol. 41, American Documentation Institute (1990), Nr. 6, p. 391.
- [DGH10] Doerr Martin, Gradmann Stefan, Hennicke Steffen, Isaac Antoine, Meghini Carlo, and van de Sompel Herbert. "The europeana data model (edm)". In: *World Library and Information Congress: 76th IFLA general conference and assembly*, 2010, pp. 10–15.
- [Dnb16] DNB. "The Integrated Authority File (GND)". URL <http://www.dnb.de/EN/gnd>. - retrieved 2017-11-21.
- [Dors17] Dorsch Isabelle. "Relative visibility of authors' publications in

- different information services". In: *Scientometrics* vol. 112, Springer Netherlands (2017), Nr. 2, pp. 917–925.
- [DSEQ13] Dietze Stefan, Sanchez-Alonso Salvador, Ebner Hannes, Qing Yu Hong, Giordano Daniela, Marenzi Ivana, and Pereira Nunes Bernardo. "Interlinking educational resources and the web of data". In: *Program* vol. 47, Emerald Group Publishing Limited (2013), Nr. 1, pp. 60–91.
- [EESB18] Elekes Ábel, Englhardt Adrian, Schäler Martin, and Böhm Klemens. "Toward meaningful notions of similarity in NLP embedding models". In: *International Journal on Digital Libraries*, Springer (2018), pp. 1–20.
- [EIIV07] Elmagarmid Ahmed K., Ipeirotis Panagiotis G., and Verykios Vassilios S. "Duplicate Record Detection: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* vol. 19 (2007), Nr. 1, pp. 1–16.
- [FCLV11] Fernández Miriam, Cantador Iván, López Vanesa, Vallet David, Castells Pablo, and Motta Enrico. "Semantically enhanced Information Retrieval: An ontology-based approach". In: *Web Semantics: Science, Services and Agents on the World Wide Web* vol. 9 (2011), Nr. 4, pp. 434–452.
- [FeGL12] Ferreira Anderson A, Gonçalves Marcos André, and Laender Alberto H F. "A brief survey of automatic methods for author name disambiguation". In: *Acm Sigmod Record* vol. 41, ACM (2012), Nr. 2, pp. 15–26.
- [FeMG12] Ferreira Anderson A, Machado Tales Mota, and Gonçalves Marcos André. "Improving author name disambiguation with user relevance feedback". In: *Journal of Information and Data Management* vol. 3 (2012), Nr. 3, pp. 332–347.
- [Fens01] Fensel Dieter. "Ontologies". In: *Ontologies* : Springer, Berlin, Heidelberg, 2001, pp. 11–18. — ISBN 978-3-662-04396-7.
- [FGMR01] Finkelstein Lev, Gabrilovich Evgeniy, Matias Yossi, Rivlin Ehud, Solan Zach, Wolfman Gadi, and Ruppin Eytan. "Placing

Bibliography

- search in context: The concept revisited". In: *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 406–414.
- [FMFG18] Fernández Javier D, Martínez-Prieto Miguel A, de la Fuente Redondo Pablo, and Gutiérrez Claudio. "Characterising RDF data sets". In: *Journal of Information Science* vol. 44 (2018), Nr. 2, pp. 203–229.
- [FTRD16] Faruqi Manaal, Tsvetkov Yulia, Rastogi Pushpendre, and Dyer Chris. "Problems With Evaluation of Word Embeddings Using Word Similarity Tasks" (2016), cs.CL/1605.02276.
- [FWMJ12] Freire Nuno, Wiermer Rene, Muhr Markus, Juffinger Andreas, and Latronico Chiara. "Author Consolidation across European National Bibliographies and Academic Digital Repositories". In: *11th International Conference on Current Research Information Systems*, (2012), pp. 203–210. — ISBN 9788086742335.
- [FWPZ11] Fan Xiaoming, Wang Jianyong, Pu Xu, Zhou Lizhu, and Lv Bing. "On graph-based name disambiguation". In: *Journal of Data and Information Quality (JDIQ)* vol. 2, ACM (2011), Nr. 2, p. 10.
- [GaGF10] Garibay Cecilia, Gutiérrez Humberto, and Figueroa Arturo. "Evaluation of a Digital Library by Means of Quality Function Deployment (QFD) and the Kano Model". In: *The Journal of Academic Librarianship* vol. 36 (2010), Nr. 2, pp. 125–132.
- [GiZH05] Giles C Lee, Zha Hongyuan, and Han Hui. "Name disambiguation in author citations using a k-way spectral clustering method". In: *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, 2005, pp. 334–343.
- [GMSB17] Galke Lukas, Mai Florian, Schelten Alan, Brunsch Dennis, and Scherp Ansgar. "Comparing Titles vs. Full-text for Multi-Label Classification of Scientific Papers and News Articles". In: *arXiv preprint arXiv:1705.05311* (2017).

- [GRMJ15] Ganguly Debasis, Roy Dwaipayan, Mitra Mandar, and Jones Gareth J F. "Word embedding based generalized language model for information retrieval". In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 795–798.
- [GTCG13] Godoi Thiago A, Torres Ricardo da S, Carvalho Ariadne M B R, Gonçalves Marcos A, Ferreira Anderson A, Fan Weiguo, and Fox Edward A. "A relevance feedback approach for the author name disambiguation problem". In: *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, 2013, pp. 209–218.
- [Haak13] Haak Laurel L. "ORCID: connecting researchers and scholars with their works". In: *Insights: the UKSG journal* vol. 26, UKSG in association with Ubiquity Press (2013), Nr. 3, pp. 239–243.
- [HaLT14] Hajra Arben, Latif Atif, and Tochtermann Klaus. "Retrieving and ranking scientific publications from linked open data repositories". In: *14th International Conference on Knowledge Technologies and Data-driven Business (i-KNOW '14)*. New York, NY, USA, ACM Press (2014), pp. 1–4 — ISBN 9781450327695.
- [HaPa17] Hartmann Sarah, and Pampel Heinz. "GND und ORCID: Brückenschlag zwischen zwei Systemen zur Autorenidentifikation". In: *Bibliotheksdienst* vol. 51, De Gruyter (2017), Nr. 7, pp. 575–588.
- [HaPT21] Hajra Arben, Pianos Tamara, and Tochtermann Klaus. "Linking Author Information: EconBiz Author Profiles". In: *Metadata and Semantics Research, MTSR 2020, CCIS 1355*. Madrid, Spain, Springer International Publishing, 2021, pp. 180–191. — ISBN 978-3-030-71902-9.
- [HaRT15] Hajra Arben, Radevski Vladimir, and Tochtermann Klaus. "Author Profile Enrichment for Cross-Linking Digital Libraries". In: *Research and Advanced Technology for Digital Libraries, TPD L*. vol. 9316, Springer, Cham, 2015, pp. 124–136 — ISBN 9783319245911.

Bibliography

- [HaSe13] Harris Steve, and Seaborne Andy. "SPARQL 1.1 query language". *W3C recommendation*, <https://www.w3.org/TR/2013/REC-sparql11-query-20130321/>, *W3C(Mar 2013)*, 2013.
- [HaTo16] Hajra Arben, and Tochtermann Klaus. "Enriching scientific publications from LOD repositories through word embeddings approach". In: *In Conf. on Metadata and Semantics Research MTSR 2016*, CCIS vol. 672, Springer, Cham, 2016, pp. 278–290. — ISBN 9783319491561.
- [HaTo17] Hajra Arben, and Tochtermann Klaus. "Linking science: approaches for linking scientific publications across different LOD repositories". In: *Int. Journal of Metadata, Semantics and Ontologies* vol. 12 (2017), Nr. 2/3, pp. 121–141.
- [HaTo18] Hajra Arben, and Tochtermann Klaus. "Visual Search in Digital Libraries and the Usage of External Terms". In: *22nd International Conference Information Visualisation (IV)*: IEEE, 2018, pp. 396–400. — ISBN 978-1-5386-7202-0.
- [HCOC02] Huang Zan, Chung Wingyan, Ong Thian-Huat, and Chen Hsinchun. "A graph-based recommender system for digital library". In: *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries - JCDL '02*. New York, NY, USA : ACM Press, 2002, p. 65. — ISBN 1581135130.
- [HCRP07] Hersh William, Cohen Aaron, Ruslen Lynn, and Phoebe Roberts. "TREC 2007 genomics track overview". In: *The Sixteenth Text Retrieval Conference (TREC 2007)* : Gaithersburg, MD: National Institute for Standards & Technology, 2007.
- [HeBi11] Heath Tom, and Bizer Christian. "Linked data: Evolving the web into a global data space". In: *Synthesis lectures on the semantic web: theory and technology* vol. 1, Morgan & Claypool Publishers (2011), Nr. 1, pp. 1–136.
- [HFCH12] Heradio R., Fernandez-Amoros D., Cabrerizo F. J., and Herrera-Viedma E. "A review of quality evaluation of digital libraries based on users' perceptions". In: *Journal of Information Science* vol. 38 (2012), Nr. 3, pp. 269–283.

- [HGZL04] Han Hui, Giles Lee, Zha Hongyuan, Li Cheng, and Tsioutsoulouklis Kostas. "Two Supervised Learning Approaches for Name Disambiguation in Author Citations". In: *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '04*. New York, NY, USA : ACM, 2004, pp. 296–305. — ISBN 1-58113-832-6.
- [Hick16] Hickey Thomas B. "VIAF Reflections", Presentation, URL <https://www.oclc.org/content/dam/oclc/events/2016/IFLA2016/presentations/VIAF-Reflections.pdf>, 2016.
- [HiKR09] Hitzler Pascal, Krotzsch Markus, and Rudolph Sebastian. "Foundations of semantic web technologies" : Chapman and Hall/CRC, 2009 — ISBN 9781420090505.
- [HiRK16] Hill Felix, Reichart Roi, and Korhonen Anna. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation". In: *Computational Linguistics*, MIT Press (2016).
- [HiTo14] Hickey Thomas B. and Toves Jenny A. "Managing ambiguity in VIAF". In: *D-Lib Magazine* vol. 20 (2014), Nr. 7/8.
- [Hora10] Horava Tony. "Challenges and Possibilities for Collection Management in a Digital Age". In: *Library Resources & Technical Services* vol. 54 (2010), Nr. 3, pp. 142–152.
- [HSMN12] Huang Eric H, Socher Richard, Manning Christopher D, and Ng Andrew Y. "Improving Word Representations via Global Context and Multiple Word Prototypes". In: *50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, vol. 1, ACL '12. Stroudsburg, PA, USA : Association for Computational Linguistics, 2012, pp. 873–882.
- [Jacs05] Jacsó Péter. "Google Scholar: the pros and the cons". In: *Online Information Review* vol. 29, Emerald Group Publishing Limited (2005), Nr. 2, pp. 208–214.
- [JäKe00] Järvelin Kalervo, and Kekäläinen Jaana. "IR evaluation methods for retrieving highly relevant documents". In: *23rd annual international ACM SIGIR conference on Research and*

Bibliography

- development in information retrieval - SIGIR '00*. New York, NY, USA : ACM Press, 2000, pp. 41–48. — ISBN 1581132263.
- [JäKe02] Järvelin Kalervo, and Kekäläinen Jaana. "Cumulated gain-based evaluation of IR techniques". In: *ACM Transactions on Information Systems* vol. 20, ACM (2002), Nr. 4, pp. 422–446.
- [JGBM16] Joulin Armand, Grave Edouard, Bojanowski Piotr, and Mikolov Tomas. "Bag of tricks for efficient text classification". In: *arXiv preprint arXiv:1607.01759* (2016).
- [JGPH05] Joachims Thorsten, Granka Laura, Pan Bing, Hembrooke Helene, and Gay Geri. "Accurately Interpreting Clickthrough Data As Implicit Feedback". In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05*. New York, NY, USA : ACM, 2005, pp. 154–161.
- [JJHY12] Joshi Amit Krishna, Jain Prateek, Hitzler Pascal, Yeh Peter Z., Verma Kunal, Sheth Amit P., and Damova Mariana. "Alignment-Based Querying of Linked Open Data". In: *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* : Springer, Berlin, Heidelberg, 2012, pp. 807–824.
- [KeRi15] Kenter Tom, and de Rijke Maarten. "Short Text Similarity with Word Embeddings". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*. New York, NY, USA : ACM Press, 2015, pp. 1411–1420. — ISBN 9781450337946.
- [KiDi16] Kim Jinseok, and Diesner Jana. "Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks". In: *Journal of the Association for Information Science and Technology* vol. 67, Wiley Online Library (2016), Nr. 6, pp. 1446–1461.
- [KiWL16] Kim Sun, Wilbur W John, and Lu Zhiyong. "Bridging the gap: a semantic similarity measure between queries and documents". In: *arXiv preprint arXiv:1608.01972* (2016).

- [KIMc99] Kling R, and McKim G. "Scholarly communication and the continuum of electronic publishing". In: *arXiv preprint cs/9903015* (1999).
- [KNLJ09] Kang In-Su, Na Seung-Hoon, Lee Seungwoo, Jung Hanmin, Kim Pyung, Sung Won-Kyung, and Lee Jong-Hyeok. "On co-authorship for author disambiguation". In: *Information Processing & Management* vol. 45, Elsevier (2009), Nr. 1, pp. 84–97.
- [KnWi15] Knijnenburg Bart P. and Willemsen Martijn C. "Evaluating recommender systems with user experiments". In: *Recommender Systems Handbook* : Springer, 2015, pp. 309–352.
- [KrZi13] Krichel Thomas, and Zimmermann Christian. "Author Identification in Economics, ... and Beyond". In: *Working Paper Series of the German Council for Social and Economic Data* vol. 222, German Council for Social and Economic Data (RatSWD) (2013).
- [KSKW15] Kusner Matt J., Sun Yu, Kolkin Nicholas I., and Weinberger Kilian Q. "From word embeddings to document distances". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, JMLR.org (2015).
- [LaBT14] Latif Atif, Borst Timo, and Tochtermann Klaus. "Exposing Data From an Open Access Repository for Economics As Linked Data". In: *D-Lib Magazine* vol. 20, Corporation for National Research Initiatives (2014), Nr. 9, p. 2.
- [LaST16] Latif Atif, Scherp Ansgar, and Tochtermann Klaus. "LOD for Library Science: Benefits of Applying Linked Open Data in the Digital Library Setting". In: *KI - Künstliche Intelligenz* vol. 30, Springer Berlin Heidelberg (2016), Nr. 2, pp. 149–157.
- [LaSw99] Lassila Ora, and Swick Ralph R. "Resource description framework (RDF) model and syntax specification", URL <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222>, 1999. — W3C Proposed Recommendation.

Bibliography

- [LeCo15] Lebrete Rémi, and Collobert Ronan. "Rehabilitation of Count-Based Models for Word Vector Representations". In: *International Conference on Intelligent Text Processing and Computational Linguistics* : Springer, Cham, 2015, pp. 417–429.
- [LeGD15] Levy Omer, Goldberg Yoav, and Dagan Ido. "Improving distributional similarity with lessons learned from word embeddings". In: *Transactions of the Association for Computational Linguistics* vol. 3 (2015), pp. 211–225.
- [LGCF08] Laender Alberto H.F., Gonçalves Marcos André, Cota Ricardo G., Ferreira Anderson A., Santos Rodrygo L. T., and Silva Allan J.C. "Keeping a digital library clean: new solutions to old problems". In: *Proceedings of the 8th ACM symposium on Document engineering*. New York, NY, USA, ACM (2008), pp. 257–262. — ISBN 978-1-60558-081-4.
- [LIKC14] Liu Wanli, Islamaj Doğan Rezarta, Kim Sun, Comeau Donald C, Kim Won, Yeganova Lana, Lu Zhiyong, and Wilbur W John. "Author Name Disambiguation for PubMed". In: *Journal of the Association for Information Science and Technology* vol. 65, NIH Public Access (2014), Nr. 4, pp. 765–781.
- [LiWi07] Lin Jimmy, and Wilbur W John. "PubMed related articles: a probabilistic topic-based model for content similarity". In: *BMC bioinformatics* vol. 8, BioMed Central (2007), Nr. 1, p. 423.
- [LLHZ16] Lai Siwei, Liu Kang, He Shizhu, and Zhao Jun. "How to generate a good word embedding". In: *IEEE Intelligent Systems* vol. 31, IEEE (2016), Nr. 6, pp. 5–14.
- [Loes11] Loesch Martha Fallahay. "VIAF (The Virtual International Authority File) – <http://viaf.org>". In: *Technical Services Quarterly* vol. 28, Routledge (2011), Nr. 2, pp. 255–256.
- [LoGS11] Lops Pasquale, de Gemmis Marco, and Semeraro Giovanni. "Content-based Recommender Systems: State of the Art and Trends". In: *Recommender Systems Handbook*. Boston, MA : Springer US, 2011, pp. 73–105.

- [LyGa96] Lynch Clifford, and Garcia-Molina Hector. "Interoperability, Scaling, and the Digital Libraries Research Agenda". In: *Microcomputers for Information Management* vol. 13, ERIC (1996), Nr. 2, pp. 85–132.
- [MaAG13] MacEwan Andrew, Angjeli Anila, and Gatenby Janifer. "The International Standard Name Identifier (ISNI): The Evolving Future of Name Authority Control". In: *Cataloging & Classification Quarterly* vol. 51, Taylor & Francis Group (2013), Nr. 1–3, pp. 55–71.
- [MaRS08] Manning Christopher D, Raghavan Prabhakar, and Schütze Hinrich. "Introduction to Information Retrieval". New York, USA: Cambridge University Press, 2008 — ISBN 0521865719, 9780521865715.
- [MaSt01] Maedche Alexander, and Staab Steffen. "Ontology learning for the semantic web". In: *IEEE Intelligent systems* vol. 16, IEEE (2001), Nr. 2, pp. 72–79.
- [MaYa03] Mann Gideon S, and Yarowsky David. "Unsupervised personal name disambiguation". In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 2003, pp. 33–40.
- [MBLG17] Musto Cataldo, Basile Pierpaolo, Lops Pasquale, de Gemmis Marco, and Semeraro Giovanni. "Introducing linked open data in graph-based recommender systems". In: *Information Processing & Management* vol. 53, Elsevier (2017), Nr. 2, pp. 405–435.
- [MCCD13] Mikolov Tomas, Chen Kai, Corrado Greg, and Dean Jeffrey. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).
- [McHO04] McGuinness Deborah L, Van Harmelen Frank, and others. "OWL web ontology language overview". URL <http://www.w3.org/TR/2004/REC-owl-features-20040210>, — W3C recommendation (2004).

Bibliography

- [MeDa15] Meymandpour Rouzbeh, and Davis Joseph G. "Enhancing recommender systems using linked open data-based semantic analysis of items". In: *Proceedings of the 3rd australasian web conference (AWC 2015)*. vol. 27, 2015, p. 30.
- [MGBP17] Mikolov Tomas, Grave Edouard, Bojanowski Piotr, Puhersch Christian, and Joulin Armand. "Advances in Pre-Training Distributed Word Representations". In: *arXiv preprint arXiv:1712.09405* (2017).
- [Mill95] Miller George A. "WordNet: a lexical database for English". In: *Communications of the ACM* vol. 38, New York, NY, USA: Commun. ACM (1995), Nr. 11, pp. 39–41.
- [MnHi09] Mnih Andriy, and Hinton Geoffrey E. "A scalable hierarchical distributed language model". In: *Advances in neural information processing systems*, 2009, pp. 1081–1088.
- [MNLG18] Musto Cataldo, Narducci Fedelucio, Lops Pasquale, de Gemmis Marco, and Semeraro Giovanni. "Linked open data-based explanations for transparent recommender systems". In: *International Journal of Human-Computer Studies*, Elsevier (2018).
- [Mood16] Moody Christopher E. "Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec". In: *arXiv preprint arXiv:1605.02019*, 2016.
- [MoRo00] Mooney Raymond J., and Roy Loriene. "Content-based book recommending using learning for text categorization". In: *5th ACM conference on Digital libraries - DL '00*. New York, NY, USA : ACM Press, 2000, pp. 195–204. — ISBN 158113231X.
- [MSGL16] Musto Cataldo, Semeraro Giovanni, de Gemmis Marco, and Lops Pasquale. "Learning word embeddings from wikipedia for content-based recommender systems". In: *European Conference on Information Retrieval*, 2016, pp. 729–734.
- [MüRR17] Müller Mark-Christoph, Reitz Florian, and Roy Nicolas. "Data sets for author name disambiguation: an empirical analysis

- and a new resource". In: *Scientometrics* vol. 111, Springer (2017), Nr. 3, pp. 1467–1500.
- [Must10] Musto Cataldo. "Enhanced vector space models for content-based recommender systems". In: *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 361–364.
- [NeTo12] Neubert Joachim, and Tochtermann Klaus. "Linked Library Data: Offering a Backbone for the Semantic Web". In: *Lukose D., Ahmad A.R., Suliman A. (eds) Knowledge Technology. KTW 2011* : Springer, Berlin, Heidelberg, 2012, pp. 37–45.
- [Neub09] Neubert Joachim. Bringing the "Thesaurus for Economics" on to the Web of Linked Data. In: *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009*. Madrid, Spain : CEUR-WS.org, 2009.
- [Neub17] Neubert Joachim. "Wikidata as a Linking Hub for Knowledge Organization Systems? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings". In: *NKOS@TPDL 2017*. Thessaloniki, Greece, 2017, pp. 14–25.
- [NGBM15] Niraula Nobal B, Gautam Dipesh, Banjade Rajendra, Maharjan Nabin, and Rus Vasile. "Combining Word Representations for Measuring Word Relatedness and Similarity". In: *Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference* (2015). — ISBN 9781577357308.
- [NiMW17] Nielsen Finnårup Finnårup, Mietchen Daniel, and Willighagen Egon. "Scholia and scientometrics with Wikidata". In: *European Semantic Web Conference* : Springer, Cham, 2017, pp. 237–259.
- [NMOR12] Di Noia Tommaso, Mirizzi Roberto, Ostuni Vito Claudio, Romito Davide, and Zanker Markus. "Linked open data to support content-based recommender systems". In: *Proceedings of the 8th International Conference on Semantic Systems - I-SEMANTICS '12*. New York, NY, USA : ACM Press, 2012, pp. 1–8. — ISBN 9781450311120.

Bibliography

- [Norv13] Norvig Peter. "English letter frequency counts: Mayzner revisited or ETAOIN SRHLCU". In: *Norvig. com* (2013).
- [Oarr17] OARR. "Open Access Repository Ranking". URL <http://repositoryranking.org/>. - retrieved 2017-02-15.
- [PaBi07] Pazzani Michael J. and Billsus Daniel. "Content-based recommendation systems". In: *The adaptive web*: Springer, 2007, pp. 325–341.
- [PaKS15] Papadakis Ioannis, Kyprianos Konstantinos, and Stefanidakis Michalis. "Linked data URIs and libraries: the story so far". In: *D-Lib Magazine* vol. 21 (2015), Nr. 5/6.
- [Pass10a] Passant Alexandre. "Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations". In: *AAAI spring symposium: linked data meets artificial intelligence*. vol. 77, 2010, p. 123.
- [Pass10b] Passant Alexandre. "dbrec – music recommendations using DBpedia". In: *International Semantic Web Conference*, 2010, pp. 209–224.
- [PCWG98] Paepcke Andreas, Chang Chen-Chuan K., Winograd Terry, and García-Molina Héctor. "Interoperability for digital libraries worldwide". In: *Communications of the ACM* vol. 41 (1998), Nr. 4, pp. 33–42.
- [PeSM14] Pennington Jeffrey, Socher Richard, and Manning Christopher D. "Glove: Global vectors for word representation". In: *EMNLP*. vol. 14, 2014, pp. 1532–1543.
- [PeVo13] Peska Ladislav, and Vojtas Peter. "Enhancing recommender system with linked open data". In: *International Conference on Flexible Query Answering Systems*, 2013, pp. 483–494.
- [PiHa21] Pianos Tamara, and Hajra Arben. "Automatisch erzeugte Author Profiles als Grundlage für eine Themenseite zur Corona-Krise". In: *ABI Technik* vol. 41, Walter de Gruyter GmbH (2021), Nr. 1, pp. 13–20.

- [PKCK12] Park Deuk Hee, Kim Hyea Kyeong, Choi Il Young, and Kim Jae Kyeong. "A literature review and classification of recommender systems research". In: *Expert Systems with Applications* vol. 39 (2012), Nr. 11, pp. 10059–10072.
- [PrSe08] Prud'hommeaux Eric, and Seaborne Andy. "SPARQL query language for RDF". URL <https://www.w3.org/TR/rdf-sparql-query/>, – W3C Recommendation (2008).
- [PRZL09] Pereira Denilson Alves, Ribeiro-Neto Berthier, Ziviani Nivio, Laender Alberto H F, Gonçalves Marcos André, and Ferreira Anderson A. "Using web information for author name disambiguation". In: *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 2009, pp. 49–58.
- [RaHL16] Radevski Vladimir, Hajra Arben, and Limani Fidan. "Semantically Related Data as Technology-Enhanced Support for Research Assistive and Quality Tools". In: *Technology Advanced Quality Learning for ALL, QED'16*, 2016, pp. 18–31.
- [Ramo03] Ramos Juan. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. vol. 242, 2003, pp. 133–142.
- [ReRe06] Recordon David, and Reed Drummond. "OpenID 2.0: A Platform for User-centric Identity Management". In: *Proceedings of the Second ACM Workshop on Digital Identity Management, DIM '06*. New York, NY, USA : ACM, 2006, pp. 11–16. – ISBN 1-59593-547-9.
- [Resn61] Resnick A. "Relative effectiveness of document titles and abstracts for determining relevance of documents". In: *Science* vol. 134, American Association for the Advancement of Science (1961), Nr. 3484, pp. 1004–1006.
- [ŘeSo10] Řehůřek Radim, and Sojka Petr. "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta : ELRA, 2010, pp. 45–50.

Bibliography

- [RiRS11] Ricci Francesco, Rokach Lior, and Shapira Bracha. "Introduction to recommender systems handbook". In: *Recommender systems handbook* : Springer, 2011, pp. 1–35.
- [RoHa17] Rowley Jennifer, and Hartley Richard. "Organizing knowledge: an introduction to managing access to information", Routledge, 4th ed., 2017. — ISBN 978-0754644316.
- [RoZa10] Robertson Stephen, and Zaragoza Hugo. "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends® in Information Retrieval* vol. 3, Now Publishers Inc. (2010), Nr. 4, pp. 333–389.
- [SaBu88] Salton Gerard, and Buckley Christopher. "Term-weighting approaches in automatic text retrieval". In: *Information processing & management* vol. 24, Elsevier (1988), Nr. 5, pp. 513–523.
- [Salo09] Salo Dorothea. "Name authority control in institutional repositories". In: *Cataloging & Classification Quarterly* vol. 47, Taylor & Francis (2009), Nr. 3–4, pp. 249–261.
- [SaMc86] Salton Gerard, and McGill Michael J. "Introduction to Modern Information Retrieval". New York, NY, USA : McGraw-Hill, Inc., 1986. — ISBN 0070544840.
- [SaWY75] Salton Gerard, Wong Anita, and Yang Chung-Shu. "A vector space model for automatic indexing". In: *Communications of the ACM* vol. 18, ACM (1975), Nr. 11, pp. 613–620.
- [Scha97] Schatz B R. "Information retrieval in digital libraries: bringing search to the net". In: *Science (New York, N.Y.)* vol. 275, American Association for the Advancement of Science (1997), Nr. 5298, pp. 327–34.
- [SGLF14] Santana Alan Filipe, Gonçalves Marcos André, Laender Alberto H F, and Ferreira Anderson. "Combining Domain-specific Heuristics for Author Name Disambiguation". In: *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '14*. Piscataway, NJ, USA : IEEE Press, 2014,

- pp. 173–182. — ISBN 978-1-4799-5569-5.
- [SGLF17] Santana Alan Filipe, Gonçalves Marcos André, Laender Alberto H. F., and Ferreira Anderson A. "Incremental author name disambiguation by exploiting domain-specific heuristics". In: *Journal of the Association for Information Science and Technology* vol. 68, John Wiley & Sons, Ltd (2017), Nr. 4, pp. 931–945.
- [SHCL07] Song Yang, Huang Jian, Councill Isaac G, Li Jia, and Giles C Lee. "Efficient Topic-based Unsupervised Name Disambiguation". In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07*. New York, NY, USA : ACM, 2007, pp. 342–351. — ISBN 978-1-59593-644-8.
- [Shet99] Sheth Amit P. "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics", Springer US (1999), pp. 5–29.
- [ShLM09] Shu Liangcai, Long Bo, and Meng Weiyi. "A Latent Topic Model for Complete Entity Resolution". In: *Proceedings of the IEEE International Conference on Data Engineering, ICDE '09*. Washington, DC, USA : IEEE Computer Society, 2009, pp. 880–891. — ISBN 978-0-7695-3545-6.
- [SLMJ15] Schnabel Tobias, Labutov Igor, Mimno David M, and Joachims Thorsten. "Evaluation methods for unsupervised word embeddings". In: *EMNLP*, 2015, pp. 298–307.
- [SmCa05] Smeaton Alan F., and Callan Jamie. "Personalisation and recommender systems in digital libraries". In: *International Journal on Digital Libraries* vol. 5, Springer Berlin Heidelberg (2005), Nr. 4, pp. 299–308.
- [SmTo09] Smalheiser Neil R, and Torvik Vetle I. "Author name disambiguation". In: *Annual review of information science and technology* vol. 43, Wiley Online Library (2009), Nr. 1, pp. 1–43.
- [Stef10] Stefan Evert. "Google web 1T 5-grams made easy (but not for the computer)". In: *Proceedings of the NAACL HLT 2010 Sixth*

Bibliography

Web as Corpus Workshop: Association for Computational Linguistics, 2010, pp. 32–40.

- [Stw17] STW (ZBW). "STW Thesaurus for Economics". URL <http://www.zbw.eu/en/stw-info/>. - retrieved 2017-09-26.
- [SuKa10] Sugiyama Kazunari, and Kan Min-Yen. "Scholarly paper recommendation via user's recent research interests". In: *10th annual joint conference on Digital libraries - JCDL '10*. New York, NY, USA : ACM Press, 2010, p. 29. — ISBN 9781450300858.
- [TBSB03] Tansley Robert, Bass Mick, Stuve David, Branschofsky Margret, Chudnov Daniel, McClellan Greg, and Smith MacKenzie. "The DSpace institutional digital repository system: current functionality". In: *2003 Joint Conference on Digital Libraries, 2003. Proceedings.*, 2003, pp. 87–97.
- [Tenn02] Tennant Roy. "MARC must die". In: *Library Journal-New York* vol. 127, Library Journal (2002), Nr. 17, pp. 26–27.
- [Tenn04] Tennant Roy. "A bibliographic metadata infrastructure for the twenty-first century". In: *Library Hi Tech* vol. 22, Emerald Group Publishing Limited (2004), Nr. 2, pp. 175–181.
- [TFWZ12] Tang Jie, Fong Alvis C M, Wang Bo, and Zhang Jing. "A unified probabilistic framework for name disambiguation in digital library". In: *IEEE Transactions on Knowledge and Data Engineering* vol. 24, IEEE (2012), Nr. 6, pp. 975–987.
- [Than14] Thanos Costantino. "The future of digital scholarship". In: *Procedia Computer Science* vol. 38, Elsevier (2014), pp. 22–27.
- [Than16] Thanos Costantino. "A Vision for Open Cyber-Scholarly Infrastructures". In: *Publications* vol. 4, Multidisciplinary Digital Publishing Institute (2016), Nr. 2, p. 13.
- [ToSm09] Torvik Vetle I, and Smalheiser Neil R. "Author name disambiguation in MEDLINE". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* vol. 3, ACM (2009), Nr. 3, p. 11.

- [TuPa10] Turney Peter D, and Pantel Patrick. "From frequency to meaning: Vector space models of semantics". In: *Journal of artificial intelligence research* vol. 37 (2010), pp. 141–188.
- [TuRB10] Turian Joseph, Ratinov Lev, and Bengio Yoshua. "Word Representations: A Simple and General Method for Semi-supervised Learning". In: *48th Annual Meeting of the Association for Computational Linguistics, ACL '10*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2010, pp. 384–394.
- [TWYZ14] Tang Duyu, Wei Furu, Yang Nan, Zhou Ming, Liu Ting, and Qin Bing. "Learning sentiment-specific word embedding for twitter sentiment classification". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. vol. 1, 2014, pp. 1555–1565.
- [VFRT16] Vagliano Iacopo, Figueroa Cristhian, Rocha Oscar Rodríguez, Torchiano Marco, Faron-Zucker Catherine, and Morisio Maurizio. "Redyal: A dynamic recommendation algorithm based on linked data". In: *3rd Workshop on New Trends in Content-Based Recommender Systems co-located with ACM Conf. on Recommender Systems (RecSys 2016)* vol. 1673, 2016.
- [VrKr14] Vrandečić Denny, and Krötzsch Markus. "Wikidata: A Free Collaborative Knowledgebase". In: *Commun. ACM* vol. 57. New York, NY, USA, ACM (2014), Nr. 10, pp. 78–85.
- [WiFe18] Wimalaratne Sara, and Fenner Martin. "D2.1 PID Resolution Services Best Practices", 2018. — DOI: 10.5281/ZENODO.1324300.
- [Wiki18] Wikidata. "Welcome to Wikidata". URL https://www.wikidata.org/wiki/Wikidata:Main_Page. - retrieved 2018-01-07.
- [WWLH13] Wang Yining, Wang Liwei, Li Yuanzhi, He Di, Liu Tie-YanLiu Tie-Yan, and Chen Wei. "A Theoretical Analysis of NDCG Ranking Measures". In: *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*, 2013.
- [Xie06] Xie Hong. "Evaluation of digital libraries: Criteria and

Bibliography

- problems from users' perspectives". In: *Library & Information Science Research* vol. 28 (2006), Nr. 3, pp. 433–452.
- [YSMB16] Ye Xin, Shen Hui, Ma Xiao, Bunescu Razvan, and Liu Chang. "From word embeddings to document similarities for improved information retrieval in software engineering". In: *Proceedings of the 38th International Conference on Software Engineering - ICSE '16*. New York, NY, USA : ACM Press, 2016, pp. 404–415. — ISBN 9781450339001.
- [Yu11] Yu Liyang. "A developer's guide to the semantic Web", Springer Heidelberg Dordrecht London New York, 2011. — ISBN 978-3-642-15969-5.
- [ZaCr16] Zamani Hamed, and Croft W Bruce. "Estimating Embedding Vectors for Queries". In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, ICTIR '16*. New York, NY, USA : ACM, 2016, pp. 123–132. — ISBN 978-1-4503-4497-5.
- [Zeli18] Zelikman Eric. "Context is Everything: Finding Meaning Statistically in Semantic Spaces". In: *arXiv preprint arXiv:1803.08493* (2018).
- [ZSCM13] Zou Will Y, Socher Richard, Cer Daniel, and Manning "Christopher D. Bilingual word embeddings for phrase-based machine translation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1393–1398.