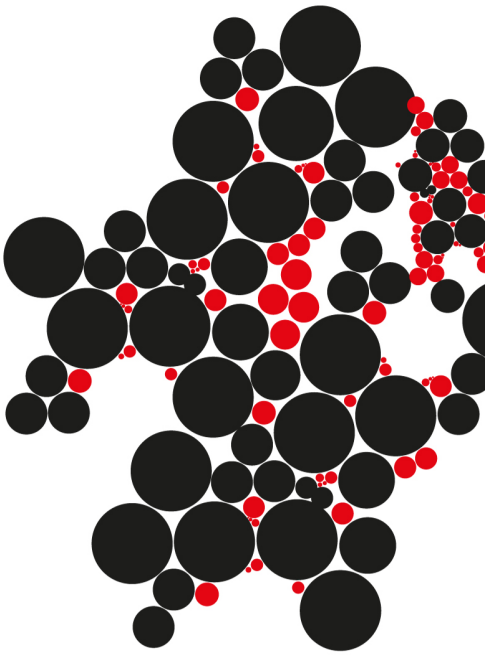
MASTER THESIS



PROBABILISTICALLY MATCHING AUTHOR NAMES TO RESEARCHERS

Ben Companjen
http://orcid.org/0000-0002-7023-9047

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE (EEMCS),
DEPARTMENT OF COMPUTER SCIENCE
DATABASES GROUP

DATA ARCHIVING AND NETWORKED SERVICES (DANS)

EXAMINATION COMMITTEE
Dr. ir. Maurice van Keulen (first supervisor)
Dr. Maarten M. Fokkinga
Ir. Maarten L. Hoogerwef (DANS)

UNIVERSITY OF TWENTE.

2013-06-20

Abstract

Publications are most important form of scientific communication, but science also consists of researchers, research projects and organisations. The goal of NARCIS (National Academic Research and Collaboration Information System) is to provide a complete and concise view of current science in the Netherlands.

Connecting publications to the researchers, projects and organisations that created them in retrospect is hard, because of a lack in the use of author identifiers in publications and researcher profiles. There is too much data to identify all researchers in NARCIS manually, so an automatic method is needed to assist completing the view of science in the Netherlands.

In this thesis the problems that limit automatic connection of author names in publications to researchers are explored and a method to automatically connect publications and researchers is developed and evaluated.

Using only the author names themselves finds the correct researcher for around 80% of the author names in an experiment, using two test sets. However, none of the correct matches were given the highest confidence of the returned matches. Over 90% of the correct matches were ranked second by confidence. Other correct matches were ranked lower, and using probabilistic results allows working with the correct results, even if they are not the best match. Many names that should not match, were included in the matches. The matching algorithm can be optimised to assign confidence to matches differently.

Including a matching function that compares publication titles and researcher's project titles did not improve the results, but better results are expected when more context elements are used to assign confidences.

Contents

1 Introduction

This chapter introduces the problem, its context, the research questions, the scope of this research, and the research method. An outline of the thesis is provided at the end of the chapter.

1.1 Motivation

"**Science** (from Latin *scientia*, meaning 'knowledge') is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe." [3]. The most important form of communication of ideas, methods and results of scientific research is the publication, text publications (articles, books, et cetera), as well as raw research data supporting conclusions. Publications can be retrieved and hence cited, which allows researchers to be recognised for their ideas and work [29]. Indexes of publications help researchers to find publications by their authors, titles, subjects and other aspects.

DANS (Data Archiving and Networked Services) is an institute of the KNAW (Royal Netherlands Academy of Arts and Sciences) and NWO (Netherlands organisation for Scientific Research). DANS's mission is to promote "sustained access to digital research data" and carries out its mission by encouraging "researchers to archive and reuse data" [10]. DANS supports the publication of, access to, preservation of and citation of research data sets through the online research data archive EASY (Electronic Archiving System). DANS supports discovery of textual and data publications through NARCIS.

NARCIS is the National Academic Research and Collaboration Information System. It is a web portal aiming to provide a transparent view of current and recent Dutch institutional science by putting key entities in context: researchers, research projects, organisations and publications (textual and data). These entity types are put in context by creating links among them, so that users can navigate from, for example, a publication via one of the authors to

research projects that the author is involved in [11]. Researchers can use NARCIS to find research and publications with underlying data related to their work, funders can track the publications following from research projects and the general public (including journalists) can find experts on certain topics.

1.2 Problem

Currently, many links are missing from the web portal, preventing the context of entities from being easily accessible via hyperlinks. NARCIS uses identifiers to link descriptions (*records*), but these identifiers need to be assigned and recorded in the records. For instance, to get from a publication to the researcher who wrote the article, the publication record needs an identifier for the researcher. The same identifier must be recorded in the researcher record to enable the portal to create a hyperlink and hence present the context to the end user. Author names are not reliable identifiers, because names are rarely unique, and a person may have multiple names or variant spellings.

The publication records in NARCIS are *harvested* from 33 institutional repositories in the Netherlands, that are managed (mostly) independently. Through national and international efforts, many of the publication records include Digital Author Identifiers (DAIs) to identify (some of) the authors. However, some of the source repositories do not support including DAIs in records and not all authors qualify to be assigned a DAI [9].

Information about researchers, research projects and organisations are retrieved from a database called VSOI (Voorloopsysteem Onderzoeksinformatie). VSOI is manually updated at DANS with data collected via nation wide yearly surveys, project overviews from funding organisations, and feedback provided by users via the NARCIS web portal.

For improving the completeness of the context of the science landscape, the options are to either try to have repositories complete the information before NARCIS harvests the records, or to try to complete the information retroactively in a best-effort automatic approach. The former is preferred, although it would need coordination among independent organisations and updating the records would take much (manual) work.

We will therefore try to solve the problem of incompleteness by taking the latter approach. In this thesis the following main research question will be addressed:

How can publications in NARCIS automatically be connected to their authors, organisations and research projects?

Standardisation efforts have resulted in the use of grant agreement numbers as identifiers for research projects [36], although they are not yet widely available in records and not all projects have grants. There are no (inter)national identifiers for organisations at the granularity of research groups. Often author affiliations are included in the publication text, but not uniformly. Names of organisations are only available in some publication records. Parsing information from publications (if the publications are even available) is outside the scope of this research.

As a first step in answering the main research question, the focus of the research is on finding the links between author name and researcher. From these links and existing links among researchers, research projects and organisations, we assume links from publications to research projects and organisations can be made.

To answer the main research question, the following four subquestions need to be answered.

1. What data is available in the Index and VSOI database and what does it look like? What (potential) problems with data quality exist?

Without identifiers for authors, finding the matching researcher has to be done by matching other attributes. This process is generally known as *entity resolution*. Names are usually not specific enough, as multiple researchers share the same name and one researcher may be known by multiple variants of a name. Therefore it is necessary to know what (other) information is available in the Index and VSOI and what the quality of the information is. How many records do include identifiers, how many researchers have links with organisations or research projects?

2. How good can the result be when just names are used?

As a baseline measurement we try to just match the author names to researchers' names. Names do not change in general, although marriage and divorce, typos, using initials or full first names etc. cause differences in spelling and possibly difficulties in matching names, but rare family names combined with five initials should result in quite certain matches.

3. What context is available? To what degree can an extended context approach improve a name only approach?

Names are included in publication records and in researcher profiles. But there is more context that may help identify the researcher whose name is on a publication, such as names of other authors and titles of projects that a researcher is or was involved with. If the context of author names and researchers are similar, the correct researcher belonging to a name can be identified more easily.

4. How do probabilistic matching results compare to non-probabilistic results?

Literature shows matching records from different record sets is often not trivial [13]. When comparing records on attribute values, the record that is most similar to the input may not be the correct match. If the best matching records are integrated or merged without further consideration, errors may be introduced.

Therefore, a model is used that allows matching of researchers to author names with alternatives. Each alternative is stored with a likelihood expressed as a probability in the range $(0, 1]$ (alternatives with probability 0 need not be stored). Even if a correct match is not the most likely as determined by the matching algorithm, the match is not lost.

Storing every possible match could also result in more work being needed afterwards, like human confirmation of correct and incorrect matches, or that so many incorrect matches are included, that the usefulness of the probabilistic result is minimal. It is therefore interesting to compare the probabilistic results to the non-probabilistic matching results.

1.3 Method

Subquestion 1 will be answered by statistical analysis of the sources. What fields and entities are available in the NARCIS Index and VSOI database? What kinds of anomalies are encountered in the records? Can the whole of the data input be used?

To match the author names to researcher profiles, a name-only approach and an extended context approach are developed. The approaches are run on a subset of the input data. The results are evaluated using both well known metrics for normal performance, used to answer subquestions 2 and 3, and probabilistic metrics to answer subquestion 4.

1.4 Outline

Background and related work are discussed in chapter 2. The available data and the statistical analysis is presented in chapter 3, in which we also explore what data is available for evaluation of the approaches. Chapter 4 describes the algorithms used to match names in the Index to researchers in VSOI. Chapter 5 describes the evaluation performed, of which the results are presented in chapter 6. Finally, chapter 7 provides conclusions of this work and recommendations for future work.

2 Background and related work

This chapter describes the theory and related research that is relevant to answering the research questions formulated in chapter 1.

Duplicate records (multiple records describing the same item) found in the NARCIS index make it more difficult to attribute publications to researchers. Therefore deduplication techniques could be useful, and they are discussed in section 2.1. Then there is the author naming problem: different authors may have very similar or even the same names. And often one author is known by multiple names or variants. Author disambiguation is the research field to find the correct real-world author for a name, which is the goal of this research too. Finally, probabilistic data integration techniques are applied in the research.

2.1 Deduplication

Semantically duplicate records in a database describe the same real world entity. In digital libraries or repositories, duplicate records may describe the same publication (e.g. article, book, conference paper) or author. When these records are in one database, the process of matching (and removing or merging) duplicate records is called deduplication [7], merge/purge [17] or entity matching, entity resolution, reference reconciliation [21]; when duplicate records need to be identified in multiple databases, this process is also called record linkage [14, 38, 7]. As explained in chapter 1, deduplication is necessary to enable identification of scientific output and creation and management of links between the correct records.

Duplicate records are created at various stages in the life of a digital library (DL). Lee et al. [23] lists these stages: creation, insertion, integration and federated search. At creation of a new digital library records that describe the same item, records need to be merged or removed except for one. After this process the newly created DL is "clean". When new records are inserted into this DL, records that would be duplicates should be removed from the inserted

records. Intuitively, duplicate records should also be handled during integration of libraries and integration of results in federated search over multiple DLs.

An overview of duplicate detection methods is given by Elmagarmid et al. [13]. Duplicate detection approaches are divided into two categories: methods that are trained to match records based on training data (including supervised and semisupervised learning and probabilistic techniques) and methods that use domain knowledge or generic distance metrics (e.g. rule based methods, record distance methods and unsupervised learning techniques) [13].

Probabilistic methods are based on probabilities that records match given (the contents of) the attributes of the records. The Bayes Decision Rules for Minimum Error or Minimum Cost can be used to draw a conclusion from the probabilities that a certain record pair belongs to the matching or nonmatching class. Some approaches also define a class of pairs that cannot be classified as matching or not matching automatically (these are in the Reject class) and need a human decision [13].

The record distance and domain knowledge based approaches use well-known string distance metrics (e.g. character based like edit distance, token based like TF.IDF, phonetics based like Soundex) to measure distance between field values on two records. Some function to convert distance (or the inverse, similarity) to a probability or decision that the record pair is a match is needed with these methods.

Panse et al. [28] describe a indeterministic approach to duplicate detection. The approach is indeterministic in the sense that record pairs are not decided to be matching or not matching. The authors experimentally derive a function to convert record similarity into duplicate probability.

A comparison of 11 frameworks for entity matching was performed by Köpcke and Rahm [21]. Most of the compared frameworks support learning of matcher combinations optimal for the matching task, blocking methods to decrease the number of needed comparisons and deduplication based on attribute values.

The Duplicate Detection toolkit (DuDe) is another framework for implementing duplicate detection algorithms [12]. Every aspect of the matching process must be configured manually, as the framework itself does not provide learning of matching parameters. Draisbach and Naumann also provide some data sets for testing with different deduplication frameworks.

PACE (Programmable Authority Control Engine) is a framework for deduplication of authority records [26]. It aims to let a collection of records (e.g.

publications or authors) be treated as authoritive records. Newly added records have to be deduplicated before they can be added. This framework is in use in the OpenAIREplus project to deduplicate records harvested from institutional repositories all over Europe.

Sometimes the same or very similar content is published in multiple forms, for example as a report, as journal article and as a conference paper. Although the contents may be the same or very similar, if the publication date or form is different, we count these as separate publications. For access to the knowledge contained in one publication, however, it may be of interest for someone to read the report version if the journal version is not easily accessible. Linking the different versions of the same content may be useful for readers that are interested in the content, but cannot access a paid version of a published article. This is not the same as deduplication, but the same techniques can be applied.

2.2 Author disambiguation

One subtask of author disambiguation can be seen as a form of deduplication because it is deduplication of records that describe the same author or references to the same author. Another subtask of disambiguation is the real disambiguation of names that refer to multiple authors. Ferreira et al. [15] summarise the distinction as follows: "(...) the same author may appear under distinct names (synonyms), or distinct authors may have similar names (polysems)." In the rest of this research this distinction is not made explicitly, although we note that both synonyms and polysems are likely to appear in the data.

The taxonomy proposed by Ferreira et al. distinguishes two author disambiguation approaches: author reference grouping and direct author assignment. The former tries to determine whether two author references refer to the same author, the latter tries to return the author record for a given author reference [15].

Author reference grouping is similar to author reference deduplication and uses many of the same methods: computing similarity of attributes or graph structure. Similarities of attributes can be computed using string similarity metrics like edit distance or token match functions applied with or without learning the best combination of similarity functions. Cohen et al. [8] suggest that a combination of Jaro-Winkler and TFIDF performs best on average when not

training the similarity function. Training similarity functions for specific problems theoretically yields better results, but it usually takes a lot of (manually edited) training data [15].

Graph structure similiarity is another form of similarity used for comparing author references [15]. A common example of such graphs is co-author graphs, in which vertices represent author names and each edge between a pair of vertices represents the co-occurence of the names in a publication. For instance, Kang et al. [19] found that co-authorship can disambiguate 85% of their test collection of authors. However, the co-author graph was extended with names from the Web, names were only in the Korean language (hence there were no synonyms) and documents with only one author were disregarded in the article.

Song et al. [32] created an approach based on topic models created from words and author references in documents. Probabilistic Latent semantic analysis and Latent Dirichlet allocation are combined with agglomerative clustering to distinguish polysem names in web pages and the Citeseer[1] database with good results: over 90% in both precision and recall.

Tang and Walsh describe three streams of disambiguation methods [33]. The first is laissez faire methods that assume random distribution of errors in names and take only exactly matching names as belonging to exactly one person. The second stream of methods acknowledges possible errors and uses fuzzy matching on author names and perhaps some other attributes, but may not work as well when information is missing between records. The third stream consists of multi-stage methods, using a simple comparator method to limit the search space and one or more methods to more exactly compare authors.

Their own method consists of clustering knowledge based on publications with shared references. Tang and Walsh assume that every researcher builds a knowledge base of literature and information learnt at conferences and that important literature is referenced in multiple publications. Often-cited publications are less important for this bibliometric profile than literature that only gets a few citations in total. If the author names in two approximately structurally equivalent publications are similar, they may refer to the same author. In two case studies accuracies of 72% and 81% were found, but a lot of time was needed to find records for all referenced articles [33].

Gurney et al. [16] propose an author grouping method based on logistic regression. It takes the attributes available in both references (among which are

[1] http://citeseerx.ist.psu.edu/

name, co-authors, common citations, important title words and others) and does not disregard records with missing values as some other methods do. It creates a graph of author references connected by weighted edges, where edge weight is the probability that two references refer to the same author. Using the clustering method from Blondel et al. [6] similar authors were clustered and identified. Average precision was above 0.9 and recall above 0.85 when using only the last name; the results were even better when using last name and first initial.

2.3 Probabilistic data integration

Normal relational databases represent a certain and complete view of the real world. Uncertain databases may represent uncertain and incomplete views of the real world, in which several statements can present alternative views on objects in the real world (Real World Objects, or RWO).

Probabilistic databases are used to store statements which are quantified by probability of, or confidence in correctness. When probabilities of alternative statements add up to 1, exactly one of the statements must be true. When the sum of probabilities of alternatives is less than 1, at most one of the statements is true. The missing probability in the incomplete distribution can be interpreted in two ways [4, 31]: the missing probability may be assigned to one unknown value or distributed over all possible tuples in the relation.

record x, name p	researcher a	0.6
record x, name p	researcher b	0.3
record x, name q	researcher c	0.5
record x, name q	researcher d	0.5

Table 2.1: Example probabilistic relation

Probabilistic data integration builds on this model by expressing uncertainty in the integration process with probabilities. Aspects of integration that may be uncertain are source data, schema mapping (e.g. table1.column1 may be similar to table2.column2 or column3), mapping of data (e.g. tuple1 may refer to the same RWO as tuple2) and queries (e.g. what is being asked, if the columns to be queried may be mapped to several other columns in the uncertain schema mapping) [25].

Applied to the domain of authors of publications and researchers in projects and organisations, probabilistic relations can specify the certainty of a tuple `name-belongs-to(author-name A, researcher B, 0.8)` with 0.8 being the confidence.

In case of an incomplete probability distribution for tuples in integration: missing probability could mean there is another unknown match that is not in the match-against set, or the correct match was inadvertently not found. That would mean that the second interpretation of missing probabilities is applicable in PDI.

3 Source data analysis

This chapter answers the first subquestion: "What data is available in the Index and VSOI database and what does it look like? What (potential) problems with data quality exist?" Statistical analysis of the NARCIS Index and VSOI data is used to determine the input parameters for applicable matching approaches as discussed in chapter 2, and to determine problems that may limit the applicability of such approaches.

3.1 Data in the NARCIS Index

The NARCIS index contains *records* of publications that a *harvesting* process collected from all Dutch universities and selected other research institutes. The widely used Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [22] is used to retrieve the records, which are formatted in the Extensible Markup Language (XML) using vocabularies like Dublin Core (DC) and Metadata Object Description Schema (MODS). An overview of the numbers of records in each format is provided in table 3.1.

Duplicate publication records exist in the Index, because sometimes an author moves from one university to another and he enters his previous publications in his new employer's repository while they are not removed from the former employer's repository. Another reason is cooperation among researchers from several universities who each deposit the work in their own reporisitory. NARCIS is an integrator of content and as such, could be a deduplicator (as described in section 2.1 and [23]). However, NARCIS does not de-duplicate the records, so it happens that a search query returns multiple records of the same publication. In some cases only the authors of the university that created the records were linked to by DAIs. For example, in a cooperation between the University of Twente and the Radboud University, the publication is deposited in the repositories of both universities. The record in the RU repository may only have DAIs for researchers they employ and the UT's record only for their

researchers. Deduplication is not trivial in these cases because of this separation of disambiguated names, as matching all authors' DAIs and title to find a match is impossible. In this research, deduplication is left to future work.

Incompleteness in records manifests itself mostly as a lack of DAIs or other identifiers. Without DAIs, the publication records are disconnected from records about researchers and without references to e.g. other publication, or projects or data sets, the context of the publication is opaque. Context is important for getting information for disambiguation from. Matching author names to identified researchers is the first step to reconstructing publication context.

Container	№of records	% of total
DIDL /MODS	706670	94.5
DIDL /DC	25716	3.44
OAI/DC	10641	1.42
RDF/OAI-ORE	4590	0.614
Total	747617	100

Table 3.1: Numbers of records in NARCIS (Index)

3.1.1 Record structure and contents

Publication records are formatted in DIDL/MODS, DIDL/DC, OAI/DC or RDF/XML.

DIDL (Digital Item Declaration Language) is a standard from the MPEG 21 group of standards for description of digital items, introduced for use in *digital libraries* by Bekaert et al. [5]. Digital library usually refers to repository systems setup to serve content (articles, books, theses, etc.) by the owner of the repository (university, but also publisher). DIDL describes the location of the objects (e.g. PDF files), access rights (closed or open access) and jump-off pages (also known as splash pages) in repositories.

MODS (Metadata Object Description Schema) is a description format from the Library of Congress that specifies a vocabulary for describing bibliographic objects [24]. Because it was developed for the American bibliographic environment, family names and given names can be put in different fields, but surname prefixes that are common in Dutch family names cannot be stored in a separate

field. Other fields include title, type (article, conference paper, book, etc.), origin (publisher, publication date, place of publication). MODS allows nesting related items inside publication records, e.g. conference proceedings with main editors nested within conference papers. Extensions can be used to store structured information that does not fit the standard MODS schema. For example, the DAI is stored in an extension of the MODS record schema.

DC (Dublin Core) is best known for its set of metadata properties. These so-called *metadata elements* can be used in a wide range of applications, but do not allow for structuring of names or relations to other objects. For example, there is only a free text field for names; no standard definition exists for including DAIs with an author name. The OAI-PMH protocol requires that DC records are supported. A simple container format (OAI/DC in table 3.1) is declared in OAI-PMH.

Either of the combinations DIDL/MODS and DIDL/DC can be used to describe publications and how to access the publications. MODS, however, is more flexible.

RDF/XML (Resource Description Format/Extensible Markup Language) is a serialisation of the RDF model and in NARCIS used to store enhanced publications. The OAI ORE (Open Archives Initiave Object Reuse and Exchange) vocabulary is used to link text publications to involved people, data sets, conference descriptions and other information. There are only about 1800 enhanced publications, which were created manually. These are not used in the process of linking authors to publications and ignored in this research.

3.1.2 Why does it look like this?

In the beginning of the national programme Digital Academic REpositories (DARE), DC was chosen as the record exchange format [34]. The DRIVER programme chose DIDL as wrapper as a solution to several problems encountered in institutional repositories, like the need for harvesting, representation of complex documents and clear use of Dublin Core identifier fields [1]. Later, most of the DC in publication records was replaced by MODS to allow the inclusion of DAIs [18].

Repository software needs to be able to export MODS and not all repositories use such software, hence not all records are available as MODS. All repository software supporting the OAI-PMH specification must support at least OAI/DC [22]; most repository software used in the Netherlands also supports MODS.

3.1.3 What can we use from records?

We need names and context to match author names to researchers. MODS records have names (separate given and family names), DAIs, and context in publication metadata. Names are associated with records with roles. 'Author' is the most common role, followed by 'editor', but in PhD theses 'thesis advisor' is also important. 'Editor' roles are mostly found in conference proceedings or books, in normal records or embedded in related items inside records for conference papers or book sections.

When looking at the names associated with a certain DAI, there are differences in spelling, use of initials or full given names, or another name (e.g. a *roepnaam*). Names without a DAI are likely to have just as many variances.

In the Netherlands, family name changes are possible when people marry (or sign a partnership contract), or when they get a divorce. It is possible that publications carry different names, and that records also carry the different names.

Name structure

The structure of names in the source data varies, because records come from different institutions that do not share a standard for entering names into institutional repositories. This lack of *name authority control* lead to the problems seen in NARCIS, and are also seen by other metadata aggregators [30].

Dublin Core records are problematic, because they do not allow for splitting names into parts or specifying whether a name is personal or corporate. Table 3.2 shows some examples of unstructured names found in Dublin Core creator and contributor fields. Because of the effort needed to identify important parts of these names, we exclude the records containing unstructured names.

Many (personal) names in MODS records are structured, i.e. split in family and given names and optionally terms of address and a name preformatted for display. Other names are unstructured: the order of name parts is unspecified and in between parentheses there may be a role description, first name or title. Names are divided into *name parts* that may be given a role (given name, family name), but sometimes the role is unspecified. Unspecified name parts could be of any type, but the largest part of these names appear to be full names in several unstructured forms. Table 3.3 shows the distribution of names by

Schoeber, J., Gemeente Venlo * Venlo (primary investigator)
Heunks, E. (RAAP)
Stevens, F. (drs.)
Sjoerd Boersma
TU Delft - CITG

Table 3.2: Examples of unstructured names in Dublin Core records

appearance of specified and unspecified name parts and DAIs. Most author names in records have a specified family name part, but 186712 (= 186698 + 11 + 3) or 8.4% have only unstructured names or only a given name. These names are disregarded too.

Family name	Given name	Unspecified	DAI	№ of names
x	x	-	-	1399183
x	x	-	x	610386
-	-	x	-	186698
x	-	-	-	14044
x	-	-	x	385
-	-	-	-	11
-	x	-	-	3

Table 3.3: Names in MODS records

Although structured names can be reordered and processed more reliably than unstructured names, there are problems in the set of structured names too, as shown in table 3.4. Apart from the names that have no family name specified (these names may have a corporate name in another field), there are some 'names' that have one full name in the family name field and another in the given name field. In a few other cases, the family name is actually a corporate name. These 'names' cannot match a person but filtering them out may not be trivial. If the matching process does try to work on these non-names, all candidate matches are incorrect, resulting in smaller parts of the results being correctly matched.

MODS uses *name parts* with *roles* to distinguish between family name and given name, but does not specify a separate role for surname prefixes. In

Data Archiving and Networked Services
A.M.B. Lips, V.J.J.M. Bekkers,
(kromhout ea)
#
-, - -

Table 3.4: Examples of strange family names in MODS records

the Index surname prefixes are entered differently in different records. The forms *<Surname, prefix>* or *<First name(s), prefix>* are common, but *<First name(s) prefix>* (without comma) and *<prefix Surname>* exist too. String matching algorithms such as the Jaro-Winkler similarity function may return different values for the same name depending on the position of the prefix. For example, "Vries, de, Jan" and "Vries, Jan, de" compare differently to "Vries, de, J.".

3.2 Data in VSOI

The VSOI database contains records about (recently) active researchers, research projects and research organisations (universities and specific other institutes with suborganisations) and the bilateral relations between pairs of these entities. It is edited manually; source data comes from an annual survey sent to universities and research institutes, collections of research information from funding organisations, suggested updates sent via the NARCIS website and information collected from news items and press releases. Research projects used to be removed a few years the end of the project, but that process was paused. Organisations are just kept up to date on a yearly basis; historic information about organisations is not available [9].

3.2.1 Record structure and contents

People are described with name (surname, surname prefix and initials), honorary title(s), expertise, external identifiers (DAI and others like identifiers from NWO, university), classifications from the NARCIS classification [2], web address and email address. If known, the person is linked to research projects that s/he participates in with a role identifier and to organisations that s/he

works for. Table 3.5 contains numbers of person, research project and organisation records with specific characteristics. People are not so connected to DAIs or other IDs, classifications or organisations. On the other hand, a majority of the people in VSOI is connected to research projects.

Entity	Number of records	%
Person	47918	100
Person with non-DAI external ID	15340	32.0
Person with DAI	20121	42.0
Person with DAI and non-DAI ID	10250	21.4
Person with involvement in project	41161	85.9
Person with classification(s)	20388	42.5
Person with expertise	16057	33.5
Person with expertise and classification(s)	15517	32.4
Person with relation to organisation	22551	47.1
Person with DAI and classification(s)	12746	26.6
Person with DAI and expertise	10477	21.9
Person with DAI, classification(s) and expertise	10312	21.5
Research	54882	100
Research with external ID	20066	36.6
Research with classification(s)	42846	78.1
Research with relation to organisation	54772	99.8
Research with relation to people	54692	99.7
Organisation	2945	100
Organisation with external ID	912	31.0
Organisation with classification(s)	2887	98.0
Organisation with relation to research	2478	84.1
Organisation with relation to people	2575	87.4

Table 3.5: Numbers of records in VSOI

In table 3.6 Person records are split by combinations of characteristics. The top 9 most occurring combinations are listed, but the listing does not reveal obvious combinations that could help match author names to researchers.

DAI	IDs	Exp.	Class	Proj	Org	№ of records	%
				x		13227	27.60
x	x	x	x	x	x	4914	10.26
x		x	x	x	x	3043	6.35
x				x		2739	5.72
	x			x		2684	5.60
		x	x	x	x	2083	4.35
				x	x	1957	4.08
x	x			x		1823	3.80
		x	x		x	1193	2.49
20121	15340	16057	20388	41161	22551	47918	
42.0%	32.0	33.5%	42.5%	85.9%	47.1%		100

Table 3.6: Numbers of people in VSOI by attribute appearance (top 9 results; 54 rows omitted)

Name structure

All names in VSOI are as complete as possible: surname (i.e. family name) and prefix, initials (no full given names), honorary title (i.e. terms of address) etc. All name parts are stored in separate fields, so the order of name parts can be determined before the names are matched.

42% of researchers have a DAI, 32% have a non-DAI identifier (e.g. university specific ID), 21.4% have both. There is no standard for including non-DAI identifiers with publications, so they cannot be reliably used. Not all researchers are assigned a DAI by the research institute and not all sources used for gathering information about researchers (e.g. research funders' websites) include DAIs [9].

3.2.2 Context

Context to add to name matching researcher names could be expertise or classifications, or organisation information, but since most people have a connection to research projects, project information should be considered. Projects have start and end dates, a title, a description, classifications and connections to organisations and other people.

Classifications are labels in NARCIS used to categorise research, organisations and researchers and allow users to filter by a topic. Using these in differentiation of author names would require that the publication metadata is somehow compared to the classification(s) found in research, organisation(s) and people. If we assume that publications are about results found in projects, then titles of projects and publications could share words.

3.3 Useful record sets

Records from the Index with DAIs that match researchers in VSOI with a DAI can be used to see how many matches were correct. On the other hand, if a DAI is not in VSOI it could mean that the name with that DAI does match any researcher. These names should not match any researcher. The names with DAIs that match or do not match researchers in VSOI will be reference sets to evaluate the performance of the match process in chapter 5. The reference sets are discussed in more detail in section 5.3.

Subset of publication records for testing: only records from 2005 up to and including 2012 are used, because information about people and research projects in VSOI is not kept forever since NARCIS is about current and recent research; information about people and research projects is removed a few years after projects finish or a person retires or dies. Only authors from this subset are used in the experiment. Reference sets are also limited to authors in the author subset.

3.4 Summary: Why don't author names and researcher names just match?

There is no *name authority control* [30] in the Netherlands: author names are not taken from an author name thesaurus that keeps a distinct name for every distinct author. Standards for sharing metadata exist (SURF, DAI, etc.), but name authority control is not among them. Hence:

- Spelling differences
- Name changes (marriage, divorce)
- Names in different formats: initials vs. full names, surname prefixes in given name/surname field or in separate field

- Different people with same name,
- Lack of DAIs, so no direct identification of who is who
- Non-names as names.

Implications: we have to work with the aggregated data, limit the scope of the solution to complete names. Family names are most important, given names may be useful, and context such as projects that researchers are involved in with (potential) co-authors can possibly help disambiguate authors with similar names.

4 Matching approaches

This chapter describes the approaches to matching author names from publications to researcher records in VSOI. The first approach, the *Name-only approach*, uses individual author names only, the second, the *Extended context approach*, adds comparisons of publication titles to project titles as context information and explores how this improves results.

4.1 General approach

Our approach to matching an author name to researchers profiles consists of calculating the similarity of the two objects, assigning a probability of a match based on that similarity and finally transforming the individual probabilities into a probability distribution for alternatives.

Without accompanying identifiers, names from publication records and researcher records from VSOI have to be matched on the contents. String comparisons (of names, project and work titles and names of co-workers and co-authors) are used for determining the similarity of an author and researcher, as we assume that in general, a higher similarity means a higher probability of matching.

The names are the most important strings to compare. Some names are so unique that they can globally identify a person; others are so common that outside a specific context it remains unclear who is meant by a name.

String similarities used in the study (Jaro-Winkler and Levenshtein edit distances and Tanimoto function, see below) return a number between 0 and 1 to represent the similarity of strings. But as some names that humans would recognise as different, still have some similarity, a translation is needed from similarity to match probability. Therefore, a function was chosen to calculate the probability of objects matching based on the similarity.

After deriving probabilities for individual matches, the probabilities have to be transformed into a probability distribution of match alternatives with a total probability of 1. For example, when an author name is G. Jansen and based on name only, there are three exact matches, each individual match probability will be 1 but with our probabilistic model only one of these (or none) can be correct. Without considering the alternative that none of these researchers matches, the probability distribution would be 1/3 for each exact match.

Because not all people are in VSOI, there is a prior probability that a person cannot (or rather: should not) be found in VSOI. Two random samples of 100 names each from the Index were manually checked for having a match in VSOI. Of the first 100, 42 names matched and of the second, 38 matched. Hence in this test, on average 40% of names matches a person in VSOI. Considering that many publications are co-authored with people who are outside the scope of VSOI (e.g. researchers working abroad and most PhD students), we assume this percentage is realistic.

In the calculation of match probabilities, the test result is taken as the general prior probability that a name matches a person in VSOI. All possible matches will account for 0.4 of all (non-)match probabilities, as formalised by equation (4.1). This normalised match probability for a combination of author name n and researcher r depends on the normal match for that combination and all other matches M_n for that same name n.

$$Pr_{match,norm}(n,r) = 0.4 \times \frac{Pr_{match}(n,r)}{\sum_{m \in M_n} Pr_{match}(n,m)} \tag{4.1}$$

In the probabilistic match model, all combinations of names and researchers are theoretically possible with a probability of 0 or higher. In practice, calculating the probabilities of all possible combinations is unfeasible and one can predict that names that are very different will have a match probability of 0 and can be discarded. A blocking method is used to reduce the number of comparisons made.

4.2 Name-only approach

In the Name-only approach, an author name on a publication record is compared to the personal name in a researcher profile and a probability of match is produced. Also, the probability that no existing record matches is produced.

4.2.1 Record similarity

The total similarity of a record match is calculated as the weighted average of three string similarity measures (Jaro-Winkler, Levenshtein and Tanimoto) applied to the full name and the full name to its first initial only in equation (4.4). The parameters subscripted with i are from the Index, the ones with subscripts j are from VSOI. Parameter n is the concatenation of surname, surname prefix and given name, in that order (e.g. "Jansen Jan"). Function fio reduces the given name to the first character (first initial only), but leaves the rest of the name intact (e.g. "Jansen J"). Function jw is the Jaro-Winkler similarity function. Function lev is the Levenshtein similarity function, which is given a low weight because the surname prefix order's impact is too big. Function $tani$ is a function based on the Tanimoto function that calculates the ratio of overlapping whitespace-separated tokens in two strings, after punctuation has been replaced by whitespace. The $tani$ function was designed to count all matching elements of a name including initials, although it does not take into account the order of name parts.

$$sim_f(n_i, n_j) = avg\left(jw(n_i, n_j), \frac{lev(n_i, n_j)}{4}, tani(n_i, n_j)\right) \tag{4.2}$$

$$sim_g(n_i, n_j) = sim_f(fio(n_i), fio(n_j)) \tag{4.3}$$

$$sim(n_i, n_j) = avg\left(sim_f(n_i, n_j), sim_g(n_i, n_j)\right) \tag{4.4}$$

4.2.2 Match probability

The probability of a match is calculated based on the similarity of a combination of family and given names. Panse et al. [28] found that it is common that similarity is mapped one to one onto match probability. Intuitively however, below a threshold record similarity (equation (4.4)) two names certainly do not match. Experiments to find a relation between similarity and match probability by Panse et al. [28] led to a $sim2p$ function, which is almost a straight line between (0.7, 0) and (1, 1). A small sample in this work's input dataset suggests that correct matches have similarities ≥ 0.6. Using a margin of 0.1 below 0.6, a simple function to derive a match probability from a similarity score is defined as (4.5).

$$Pr_{match}(n_i, n_j) = \begin{cases} (sim(n_i, n_j) - 0.5)/0.5 & 0.5 \leq sim(n_i, n_j) \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

4.2.3 Matches left out

Both sets of names were sorted in alphabetical order on last name and [with a margin of 3 on the Levenshtein edit distance around the names to compare in the context of the sorted names], names from the Index were compared to names in VSOI. Instead of $m \times n$ comparisons, the number of comparisons (and thus maximal number of resulting matches) is $O(m + n)$.

Because of bad names that do not sort correctly (e.g. surname prefixes at the beginning of last names or names with typos in the first letters), some possibly correct combinations may not be considered. For example, had there been a surname "van der Berg", it would not have been compared to "Berg" as last name with "van der" as surname prefix.

Names that came before "Aa" according to MySQL's utf-8 collation were not taken into account in the comparisons. These 225 names included non-names like "-, - -", "-LiisaHartikainen" and "#", but also 31 instances of "A, van der". The latter appears to be a real last name. However, excluding this one name will not affect the results much.

Candidate matches with a probability of 0 will be discarded.

4.3 Extended context approach

In the Extended context approach not only name similarity counts, but similarity of publication title and titles of projects the person is involved in count as well.

As noted in chapter 3, 85.9% of the researchers in VSOI are involved with at least one project. Under the assumption that titles of publications of an author have words in common with the titles of projects that the researcher works or worked on, comparing the words in lists of projects titles to the publication title increases similarity. Dissimilarity is harder, because not all project titles are available and people may change research focus not matching project titles in VSOI.

4.3.1 Record similarity

In the extended approach, project titles are the most important extra weight in the similarity score. When an author matches a researcher with project involvement(s), the Tanimoto score of the publication title and every project title is calculated. For every project title the Tanimoto score is calculated, and the maximum of all scores (in a publication-project combination) is used to calculate the match probability, as defined in equation (4.6).

Parameter pub_n is the publication record that contains the author name n. Set P_r is the set of project records that researcher r is connected to. Function *title* returns the title on a publication or project record.

$$sim_T(pub_n, P_r) = max\left(\{tani(title(pub_n), title(p)) \mid p \in P_r\}\right) \tag{4.6}$$

In equation (4.7), $sim(n_i, n_{j1})$ is the name matching score function from section 4.2.1, $sim_T(pub_n, P_r)$ is the title matching score function and $sim_{ext}(n_i, n_{j1})$ is the combined similarity score function for the extended context approach.

$$sim_{ext}(n_i, n_{j1}) = \begin{cases} 0.95 + sim_T(pub_n, P_{j1}) \times 0.05 \\ \quad \text{if } \exists j2 \wedge j1 \neq j2 \wedge sim(n_i, n_{j1}) = 1 = sim(n_i, n_{j2}) \\ sim(n_i, n_{j1}) + sim_T(pub_n, P_{j1}) \times (1 - sim(n_i, n_{j1})) \\ \quad \text{otherwise} \end{cases} \tag{4.7}$$

This function differentiates the similarities with scores from the title match process. If multiple name matches are perfect (score 1), these need to be transformed to allow the sum to be in the $[0, 1]$ range. Scaling them to make them fall in that range is not what we want, because of the incompleteness of the data - people may be involved in projects that are not in VSOI, so matches without title score are not less correct per se, To differentiate among several matches with perfect scores, a small portion of the similarity score is subtracted so that it is possible to create scores in the $[0, 1]$ range by multiplying the leftover score space.

4.3.2 Match probability

The match probability increases when titles of projects are more similar to publication titles. The same transformation function as used in the Name-only approach (but $sim(n_i, n_j)$ is replaced by $sim_{ext}(n_i, n_j)$) is used here to normalise the probabilities with one another and distribute them in the $[0, 0.4]$ range.

4.3.3 "Matches left out"

As input for the Extended context approach the same match candidates that came out of the blocking method were considered as in the Name-only approach. After processing, candidate matches with match confidence 0 are discarded.

5 Experimental setup

This chapter describes the experiment and evaluation metrics that test the performance of the approaches of chapter 4, by applying them to a subset of the data described in chapter 3 using two reference sets of known correct results.

5.1 Experiment

To answer research subquestions 2 and 3, we test the effectiveness of matching author names to the correct researchers by the Name-only and Extended context approaches of chapter 4 in an experiment. The two approaches are applied to a subset of the NARCIS Index and the (complete) VSOI data set and the results are measured using three metrics. In the subset of the Index are the author names found in publications in the years between 2005 and 2012 to decrease the time needed for the experiment. The records in the chosen subset are from recent years, because it is more likely that researchers from this period are in VSOI. The VSOI database, as the target data set, was not reduced to a subset.

The evaluation of the matching results is based on the metrics expected precision and expected recall, E_{100}(recall), precision and recall, mean reciprocal rank for (i) all matches, (ii) top 5 and (iii) top 1 matches, where applicable.

Table 5.1 describes the subsets that were used in the experiment by the numbers of records, names and DAIs used in the experiment.

5.2 Metrics

The experiment should answer the subquestions posed in chapter 1. In the evaluation, we will therefore measure the following aspects of the results.

	№ of occurrences
Publication records	276,394
All names	1,010,848
Input names	930,647
Matches in EEMCS reference set	9,851
Distinct full names / last names in EEMCS reference set	486 / 183
Distinct researchers in EEMCS reference set	154
Matches in All DAIs reference set	265,884
Distinct full names / last names in All DAIs reference set	17,867 / 13,625
Distinct researchers in All DAIs reference set	12,990
Names with DAIs not in VSOI	98,027
Distinct full names / last names	20,401 / 12,329
Distinct DAIs not in VSOI	17,164

Table 5.1: Numbers of items in the subsets used in the experiment.

How many best matches are correct? This is expressed in the precision of the positive reference sets.
How many matches should not have matched? This is expressed in the precision of the negative reference set.
How many of the known correct matches are best matched? This is expressed in the recall (at 1) of the positive reference sets.
How many of the known non-matches are not matched? This is expressed in the recall (at 1) of the negative reference set.
How many found non-matches are correctly non-matched? This is expressed in the the precision of the negative reference set.
Precision and recall are discussed in section 5.2.2.

What are the expected values of the positive matches? This is expressed in the expected precision/recall of the positive reference sets.
What are the expected values of the negative matches? This is expressed in the expected precision/recall of the negative reference set.
Expected precision and recall are discussed in section 5.2.3.

How many of the known correct (non-)matches are returned at all? This is expressed in the E_{100} recall, discussed in section 5.2.4.

Do correct matches get the highest probabilities? This is expressed in the mean reciprocal rank of correct matches, discussed in section 5.2.5.

5.2.1 Terminology

In this section, the following terms are used to explain the what the metrics measure:

(author) name (reference to a) name of an author or otherwise related person, e.g. editor, thesis advisor in a publication record

researcher record in VSOI about a researcher

null researcher a placeholder for any researcher not in the VSOI database. See also non-match.

match suggested combination (as result of running the approach) of author name and researcher

answer set of all matches for an author name

correct match a match that is correct according to a reference set

incorrect match a match that is incorrect, because the reference set shows the name in the match has a different correct match

best match match in an answer with the highest probability

non-match match involving the null researcher, i.e. the possibility that no researcher in VSOI matches the author name

Using these definitions, the result of the matching approaches can be viewed as a set of answers. An answer may include a correct match, but answers do not necessarily have one, as no match may be found. Every answer, by design of the algorithm, has a non-match.

5.2.2 Precision and recall

Precision and recall are well-known metrics in information retrieval and classification experiments. In the general case, they are used to respectively measure the ratio of correct answers in all found answers and the ratio of found correct answers in all correct answers. A high precision means most returned answers are correct (i.e. relevant to the query or correctly classified), a high recall means most of the correct answers are returned (i.e. most relevant of the relevant answers are returned, or most objects belonging to a certain class were classified as that class).

In formulas, precision and recall are defined as follows [35], in which A is the set of all found matches, G is the reference set of correct matches (ground truth) and C is the set of found matches that are correct, i.e. $C = A \cap G$:

$$Precision = \frac{|C|}{|A|} \tag{5.1}$$

$$Recall = \frac{|C|}{|G|} \tag{5.2}$$

In the experiment, true and false positive matches are counted by comparing the found matches to two reference sets of positives and true and false negative matches are counted by comparing the found matches to a reference set of negatives. The reference sets are described in detail in section 5.3. This means in one calculation of precision or recall, either true and false positives or true and false negatives can be determined.

If C of equations (5.1) and (5.2) is the set of true positives, A is the union of the sets of true positives and false positives. G is the union of the sets of true positives and false negatives. If C is the set of true negatives, A is the union of the sets of true negatives and false negatives. G is the union of the sets of true negatives and false positives.

Because the VSOI database contains a unique record for each researcher, the reference sets have only one correct match per name and only one match per name can count in the calculation of precision and recall. The best match is the obvious choice here.

If A is the set of best matches, G_+ is the reference set of correct positive matches (G_- the reference set of non-matches, i.e. matches with the null researcher) and $names(A)$ is a function to select the distinct author names in a set of matches A, then the definition of precision is as follows (where G is either G_+ or G_-, D is a subset of matches A that contain the names in the intersection of A and G):

$$Precision(A, G) = \frac{|A \cap G|}{|\{a \mid a \in D \subseteq A \wedge names(D) = names(A) \cap names(G)\}|} \tag{5.3}$$

A contains many more results that cannot be confirmed or disproved to be correct (if the reference set is a function that, given an author name from its

domain, returns a correct researcher, the domain of the function is smaller than A). In the formula, the number of correct matches is therefore divided by the number of matches in the answer of which the name is in the reference set. Similarly and using the same definitions of A and G is again either G_+ or G_-, recall is defined the same as before:

$$Recall(A, G) = \frac{|A \cap G|}{|G|} \tag{5.4}$$

5.2.3 Expected precision and recall

Expected precision extends precision by accounting for the probability that an answer is correct. If the one correct answer has a high probability and many false answers have low probability, the complete set of answers can be considered more correct than when correct answers have a low probability and false answers have high probability.

Expected recall similarly measures the probabilities of corrects answers with respect to the maximal possible probability. It is the sum of the probabilities of correct answers (which is between 0 and the number of names) divided by the total number of names.

Expected precision and recall are defined as follows by Van Keulen and De Keijzer [35] (parametrised, but same set input definitions apply; H is the answer a human would give, i.e. the reference answer):

$$E\left(Precision(C, A)\right) = \frac{E(|C|)}{E(|A|)} \tag{5.5}$$

$$E\left(Recall(C, H)\right) = \frac{E(|C|)}{|H|} \tag{5.6}$$

Parametrised with A and G (where G replaces H as the reference set, G for *ground truth*) this becomes:

$$E(Precision\left(A, G\right)) = \frac{\sum_{a \in A \cap G} Pr(a)}{\sum_{a \in A} Pr(a)} \tag{5.7}$$

$$E(Recall\,(A,G)) = \frac{\sum_{a \in A \cap G} Pr(a)}{|G|} \tag{5.8}$$

Applied using the separate reference sets, the expected precision can be defined as follows.

- $E(Precision_+)$ expected precision of positive matches
- A_+ the set of returned matches for names in the positive reference set
- G_+ the positive reference set
- C_+ the intersection of A_+ and G_+, i.e. the correct matches
- $E(Precision_-)$ expected precision of negative matches
- A_- the set of returned non-matches for names in the negative reference set
- G_- the negative reference set
- C_- the intersection of A_- and G_-, i.e. the correct non-matches

When the A, C sets are defined as sets of answers (sets of matches grouped by author name) instead of sets of independent matches and predicates for probability is specified as sum of probabilities of matches with that specific correctness, expected precision is calculated as follows:

$$E\,(Precision_+(C,A)) = \frac{\sum_{a \in C_+} Pr_{correct}(a)}{\sum_{a \in A_+} (Pr_{correct}(a) + Pr_{incorrect}(a) + Pr_{non-match}(a))} \tag{5.9}$$

$$E\,(Precision_-(C,A)) = \frac{\sum_{a \in C_-} Pr_{non-match}(a)}{\sum_{a \in A_-} (Pr_{non-match}(a) + Pr_{incorrect}(a))} \tag{5.10}$$

From the above definitions it follows that the result of expected precision is the same as that of expected recall: $\sum_{a \in A_+} (Pr_{correct}(a) + Pr_{incorrect}(a) + Pr_{non-match}(a))$ equals the number of names in C_+, since for all names $a \in A$ the probabilities of the correct match, any incorrect matches and the non-match add up to 1.

For expected precision we only consider the answers belonging to names from one reference set at a time, because the matches with other names cannot be said to be correct or incorrect. The positive reference sets are unrelated to the negative reference set and the sizes of the various reference sets is rather different. Computing a overall precision that combines positives and negatives from all reference sets is therefore not possible.

5.2.4 E_{100} recall

The E_{100} recall was introduced by Kuperus [20] to calculate how many correct results were found, regardless of the assigned probabilities to these results. In normal expected recall, if the probabilities in the answers are low (no matter how many results were correctly retrieved), the total result is low too.

E_{100} is defined in (5.11).

$$E_{100}\left(Recall\right) = \frac{|C_{all}|}{|G|} \tag{5.11}$$

where C_{all} is the set of all correct matches in the result of the matching approach (with confidence > 0) and G is the reference set (also called ground truth) of all correct matches.

This definition is similar to the definition of normal recall, but in normal recall C is different in that it contains for each name only the result with the highest probability.

To see how many correct answers are in the k matches with highest probabilities for each name, E_{100} recall at k is calculated. This is the same as equation (5.11), except that C contains only the top k answers.

It could be useful in an environment in which the top k possible answers are shown to a user who manually chooses the correct match. If the correct match is among the k answers with highest probabilities in many cases, the result can be said to be good enough and the final judgment of correctness could be manual [35]. Otherwise, other algorithms may be developed to further process the results until they are good enough.

5.2.5 Mean reciprocal rank

The mean reciprocal rank shows the average rank of the correct match in an answer (the set of matches for a name). Translated to the results in this experiment, MRR can be applied when the possible matches are ranked by probability. In (5.13) [37], A is again the set of answers, in which the matches are ordered by probability in descending order. For each answer $a \in A$, the reciprocal rank, defined in equation (5.12), is calculated. When no correct match was found, $recrank(a)$ is 0.

$$recrank(a, G) = \begin{cases} \frac{1}{rank(m)} & \text{if } \exists m : m \in a \wedge m \in G \\ 0 & \text{otherwise} \end{cases} \tag{5.12}$$

$$MRR = \frac{1}{|A|} \sum_{a=1}^{|A|} recrank(a) \tag{5.13}$$

5.3 Reference sets

Two reference sets of matches are used to evaluate the matches by the matching approach. The *positive sets* consist of names with DAIs that certainly belong to specific people in VSOI. The *negative set* contains names with DAIs that are not in VSOI.

5.3.1 Positive sets

The first positive reference set is derived from DAIs verified to be assigned correctly by the editors of the EPrints server at the Faculty of EEMCS at the University of Twente. These records were selected, because we consider the quality control process and assume the set is correct. This set contains names of people who work(ed) for the faculty.

The subset contains 9851 names, matching 154 people. There are 183 distinct family names in the set, and 486 combinations of family names and given names.

The second positive reference set is the set of all names with a DAI. This set is much larger than the EEMCS reference set: 458,817 names connected to researchers by DAI. These include the names from the EEMCS set. The quality

of the data is assumed to be slightly lower, because many different people are responsible for entering the correct DAIs and they do so using different systems and protocols [9].

The subset contains 265884 names of 12990 people. There are 13625 distinct family names, and 17867 distinct full names.

In the process of constructing the second set, some problems were found in both the publication records and the assignment of DAIs to researchers in VSOI. Some of the publication records contain malformed DAIs or author identifiers from other domains, marked up as DAI from the DAI domain. Even if the intented DAI had been correct and assigned to a researcher record, these will not match. The assignments of some DAIs in VSOI is more problematic, as the specific DAIs were linked to two records each. Table 5.2 lists the records associated with these DAIs. If these DAIs were included in the reference set, 194 names in publications would have had two correct matches. These records were not included in the reference set.

DAI	Family name	Prefix	Initials	Title	Expertise
096990090	Boonstra		A.	Dr.	Lung diseases
096990090	Boonstra		A.	Prof.dr.	ICT Strategy
158965450	Mol		E.M.M.	Dr.	
158965450	Tiecke		H.G.J.M.	Dr.	
175649847	Winter		Y.	Dr.	Formal Semantics
175649847	Vinter Seggev		Y.S.	Dr.	Formal Semantics
262496054	Faye-Visser		S.M.	Dr.	
262496054	Visser		S.M.	Dr.	
270970274	Wit	de	J.	Dr.	
270970274	Wit	de	J.	Dr.	

Table 5.2: Incorrect entries in the All DAIs reference set: the same DAI is connected to two records for obviously different people or possibly the same person with multiple records. These records were removed from the reference set.

Although these errors are serious, these are assumed to be the most important errors. Any other errors that cannot be detected are assumed to be a small fraction of the reference set and therefore their influence on the performance is assumed insignificant.

5.3.2 Negative set

The negative set was created by selecting the names in publications that have an associated DAI that is not assigned to any researcher in VSOI. If the assumption that for (nearly) all researchers who have a DAI, the DAI is registered in VSOI, is true, then the names in this reference set should not match researcher records. We should be careful with drawing conclusions, because it is hard to tell whether the assumption is correct.

If the names in the negative set do match researchers, there may be researchers in VSOI who have not been assigned their DAI. Alternatively, if the average non-match precision for the names is not significantly different from the default non-match probability assignment, we will have to be careful when drawing conclusions.

There are 98027 names in this set, distributed over 17164 DAIs. Individual DAIs in the set are found in 1 up to 142 publication records. A DAI with many occurrences in publication records that is not connected to a record, may belong to a researcher whose record was deleted (e.g. retired or deceased researchers) or whose record was not created yet (e.g. researcher may be out of scope for VSOI and NARCIS), or the DAI may not have been connected to the correct record. If an author's DAI is not registered in VSOI, but the researcher is, the precision of the negative references is lower than it should be. That is acceptable, as it means that the precision and recall calculated in the evaluation are lower bounds.

6 Evaluation

This chapter lists results from the experiment from chapter 5, evaluates the results using the listed measures and discusses how the results may be improved.

6.1 Main results of experiment

The experiment started with 930,647 names from almost 276,394 publication records, as listed in table 5.1. Of these names, 585,925 were positively matched in 5,611,824 matches; each of these matches the confidence in the correctness is more than 0. Table 6.1 shows that more than 1 million more matches were found, but with confidence 0. Hence these matches will not be used in further calculations.

Thing	№ of things
Output similarity results	7,153,291
Output similarity results (confidence > 0)	5,611,824
Output names in results	670,482
Output names in results (confidence > 0)	585,925

Table 6.1: Numbers in the experiment as the result of the Name-only approach.

The matches resulting from the experiment were regarded as answers (i.e. as sets of matches grouped by input name). Only answers for names from the reference sets can be verified. Numbers used in the evaluation metrics are in tables 6.2 (EEMCS), 6.3 (All DAIs) and 6.4 (negatives).
The values of the precision and recall calculations are listed in table 6.5; the mean reciprocal ranks are in table 6.6.

Regarding the results for names in the EEMCS and All DAIs reference sets:

	EEMCS (N)	EEMCS (E)
Matches in set	9,851	9,851
Names matched (correct or incorrect)	8,007	8,007
True positives	7,983	7,983
True positives at rank 1	0	0
True positives at rank 2	7,757	7,757
True positives at ranks 2 - 4	7,936	7,920
True positives at ranks 2 - 6	7,959	7,950
False negatives: no match at all	1,844	1,844
False positives: no correct match	24	24
Expected value of true positives	2,126.89	2,089.60
Expected value of false positives/negatives	7,724.11	7,761.40

Table 6.2: Numbers for the EEMCS reference set in the experiment as the result of the Name-only (N) and Extended context (E) approach. Ranks are based on descending order of confidence of matches within answers (equal confidences have different ranks).

	All DAIs (N)	All DAIs (E)
Matches in set	265,884	265,884
Names matched (correct or incorrect)	224,363	224,363
True positives	218,236	218,236
True positives at rank 1	0	0
True positives at rank 2	208,043	209,611
True positives at ranks 2 - 4	215,365	215,838
True positives at ranks 2 - 6	216,695	216,941
False negatives: no match at all	41,521	41,521
False positives: no correct match	6,127	6,127
Expected value of true positives	53,863.06	52,729.18
Expected value of false positives/negatives	212,020.94	213,154.82

Table 6.3: Numbers for the All DAIs reference set in the experiment as the result of the Name-only (N) and Extended context (E) approach. Ranks are based on descending order of confidence of matches within answers (equal confidences have different ranks).

- 18.7% (1,844) of the names in EEMCS and 15.6% (41,521) of the names in the All DAIs set had no match at all. This is a lot - the blocking mechanism may have been too strict (fixed name order matching could have prevented names with different name orders to match) or the similarity to probability conversion should perhaps have had a different cutoff. 24 names (0.244%) in the EEMCS (All DAIs: 6,127; 0.230%) did match, but only the wrong researchers. Because the Extended context approach did not create extra matches, these numbers did not change.
- For 78.7% of the names in EEMCS (78.2% for All DAIs) the correct match was at rank 2 (behind the non-match possibility). This equals 96.9% of the researchers in EEMCS (92.7% in All DAIs) at rank 2. The extra context information has little influence on the number of correct matches at rank 2: it decreases a little for the EEMCS set and goes up a little for the All DAIs set.
- The mean reciprocal rank agrees: average rank 2.5, which is close to the average of 78.7% at rank 2 (counting as 1/2) and 18.7% not found (counting as 0).
- Of all the found correct matches, 80.8% (EEMCS) or 81.5% (All DAIs) were ranked 6 or higher.
- The addition of title match scores in the Extended context approach did not help differentiating possible matches; scores did not change much, because the title match scores were mostly very low. Hence so the evaluation scores did not change much either between the Name-only and Extended context approach. We explore this in more detail in section 6.2.2.

Regarding the results for the names in the negatives reference set:

- For every name in the experiment, the best match is the non-match, so for the names in the negative reference set, 100% of the matches at rank 1 are correct.
- Only 33,710 names (34.4%) did not match a researcher at all. The other names matched at least one researcher, which means the names are similar enough to suggest a match.
- There is no difference at all between the results of the Name-only approach and the results of the Extended context approach. It can be that the researchers that did match, have no connections to projects, or only have connections to projects that have no words in common with the publication titles.

	Negatives (N)	Negatives (E)
Matches in set	98,027	98,027
At least one positive match	64,317	64,317
True negatives at rank 1	98,027	98,027
True negatives: no match at all	33,710	33,710
False positives: rank 2 is researcher	64,317	64,317
Expected value of true negatives	72,300.20	72,300.20
Expected value of false positives	25,726.80	25,726.80

Table 6.4: Numbers for the negative reference set in the experiment as the result of the Name-only (N) and Extended context (E) approach. Ranks are based on descending order of confidence of matches within answers (equal confidences have different ranks).

- The expected value of the true negatives are very high, which follows directly from the default assignment of confidence of 0.6 to the non-match and which is 1 when no match is found.

From the design of the algorithms defined in the Name-only approach and Extended context approach, especially the normalisation of probabilities of several positive matches and the always possible non-match, it is clear that for every name, the non-match has probability of 0.6 and the sum of probabilities for positive matches does not exceed 0.4.

This entails that the confidence given to a correct positive match can never be higher than 0.4. When multple matches are found for a name, the probabilities go down. In section 6.2.1 we will look at the distribution of confidences assigned to correct matches. Correct negatives matches (non-matches) get a relatively high confidence, because of the 0.6 non-match probability default, which becomes 1 when no match is found at all.

Expected precision and expected recall are the same, because the sum of confidences in each answer is 1, as noted in section 5.2.3. They are rather low, although it can be explained by the large influence of the prior probability for non-match. If all correct matches were found, and only correct matches were found and given probabilities of 0.40, the expected precision and recall would still have only been 0.40.

Normal precision and recall at 1 are 0 for the positive sets, because the non-match probability is always higher than any other match probability, whereas

	Prec (N)	Rec (N)	Prec (E)	Rec (E)
Normal rank 1 (EEMCS)	0	0	0	0
Normal rank 1 (all DAIs)	0	0	0	0
Normal rank 1 (non-matches)	1	1	1	1
Normal rank 2 (EEMCS)	0.969	0.787	0.969	0.787
Normal rank 2 (all DAIs)	0.927	0.782	0.934	0.788
Normal rank 2 (non-matches)	1	0.344	1	0.344
Expected (EEMCS)	0.216	0.216	0.212	0.212
Expected (all DAIs)	0.203	0.203	0.198	0.198
Expected (non-matches)	0.738	0.738	0.738	0.738
E_{100} (EEMCS)	-	0.810	-	0.810
E_{100} (all DAIs)	-	0.821	-	0.821
E_{100} (non-matches)	-	1	-	1
E_{100} at 6 (EEMCS)	-	0.808	-	0.807
E_{100} at 6 (all DAIs)	-	0.815	-	0.816
E_{100} at 6 (non-matches)	-	1	-	1

Table 6.5: Overall precision (Prec) and recall (Rec) results for Name-only (N) and Extended context (E) approach.

the names in the positive reference sets are known to have a matching researcher. On the other hand, the precision and recall of the negative reference set are 1, because for all names that are known to not match, the non-match probability is always highest.

If the matches for each answer are ranked by confidence in descending order, the non-match will always be at rank 1. If the matching approach found a best match, it will be at rank 2. It is therefore more interesting to apply the precision and recall metrics to the matches at rank 2. We then see that for the author names in the positive reference sets, 96.9% and 92.7% of the matches at rank 2 are correct, and this is even a little better in the Extended context approach. These matches make up about 78% of the correct matches. Normal rank at 2 for non-matches is the ratio of names for which no match was found at all to the total number of names.

The E_{100} recall does not change between the Name-only and Extended context approaches. By definition (see section 5.2.4), this is a measure of how many correct results were found, regardless of their probabilities. The Extended context approach starts with the results of the Name-only approach, no correct answers are discarded and no new correct answers are found.

	Name-only	Extended context
All results (EEMCS)	0.400	0.400
All results (All DAIs)	0.401	0.403
Top 5 (EEMCS)	0.400	0.399
Top 5 (All DAIs)	0.401	0.402

Table 6.6: Mean reciprocal rank for Name-only and Extended context approach.

The values of the mean reciprocal rank calculations have been summarised in table 6.6.

The matches in each answer were ranked by confidence, but multiple matches with the same confidence were ranked in random order. It is likely that for individual names this construct results in higher ranks for incorrect matches and lower ranks for correct matches, but we assume that on average this levels itself.

For the Name-only approach, the MRR is 0.401, so on average correct matches are rank 2.49. After the non-match (with probability 0.6) many correct positive

matches rank second or third. This is supported by the precision and recall of the second best match. In the Extended context approach, the MRR is 0.400. On average, the rank of the correct match is now 2.50. In the smaller set of EEMCS results, the MRR has gone down a tiny bit, although in the bigger set of results with all DAIs included, the average rank of the correct match is now 2.48. These differences are insignificantly small.

6.2 Related results

In this section, we will look at some details related to results in the previous section: the distribution of probabilities assigned to correct matches, and match scores for project and publication titles.

6.2.1 Distribution of match confidences

Figures 6.1 and 6.2 show the density of all normalised confidences for matches in the EEMCS reference set and All DAIs reference set, respectively. Density graphs show the relative distribution on a continuous domain and the total surface of the graph is 1.

The two graphs show that in most cases, the confidence assigned to the correct match after normalisation, is either 0 (or very close to 0) or 0.4 (or very close to 0.4). In the (much) larger All DAIs reference set the extreme match confidences 0 (correct match not found) and 0.4 (correct match is a perfect match) is even more clear. The blue lines show that the the confidences of correct matches in the Name-only approach concentrate around higher confidences, supporting the observations of higher expected precision and recall of the Name-only approach compared to the Extended context approach.

6.2.2 Impact of title matching

Figure 6.3 combines density curves of Tanimoto match scores, the result of equation (4.6). It shows that the great majority of project titles associated with match candidates do not match the publication title associated with the author name. A very small peak at 1 shows there are a few perfect matches. The curves are very similar, which means Dutch titles and English titles match equally bad, although Dutch project titles have more 0 scores (i.e. do not

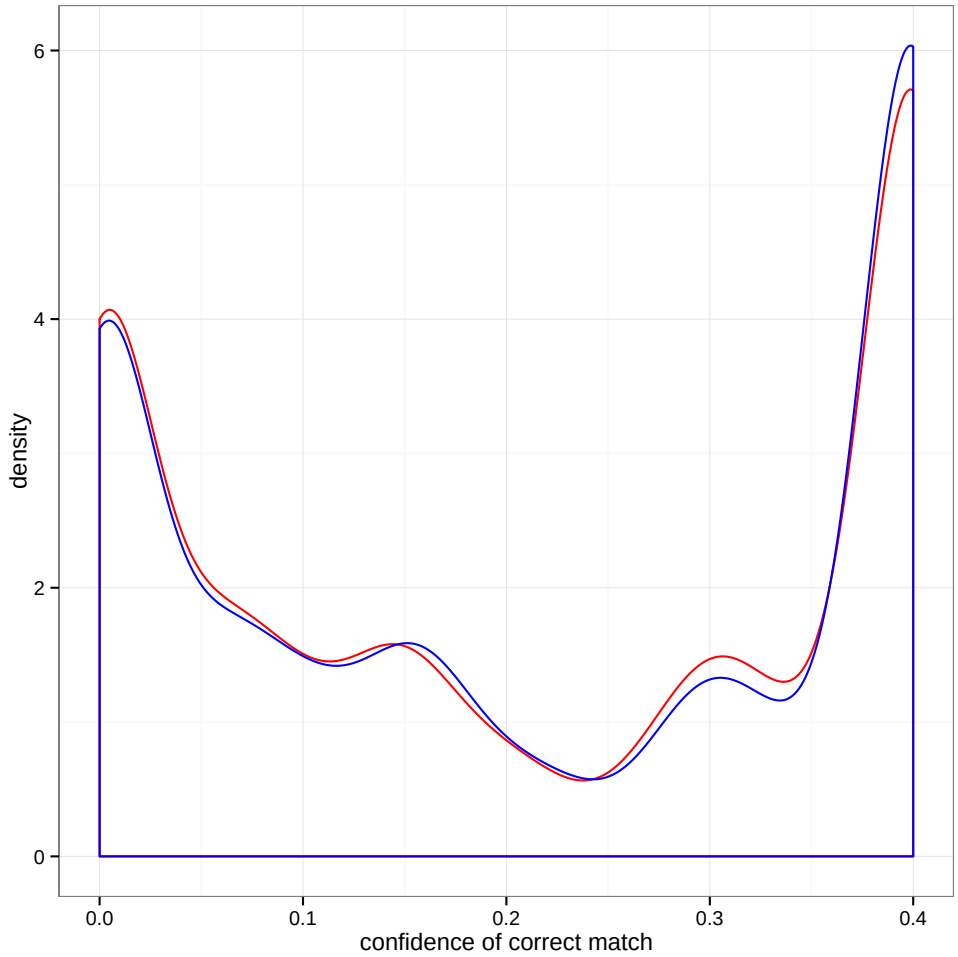


Figure 6.1: Density of confidence in correct matches in the EEMCS reference set for the Name-only approach (blue) and the Extended context approach (red).

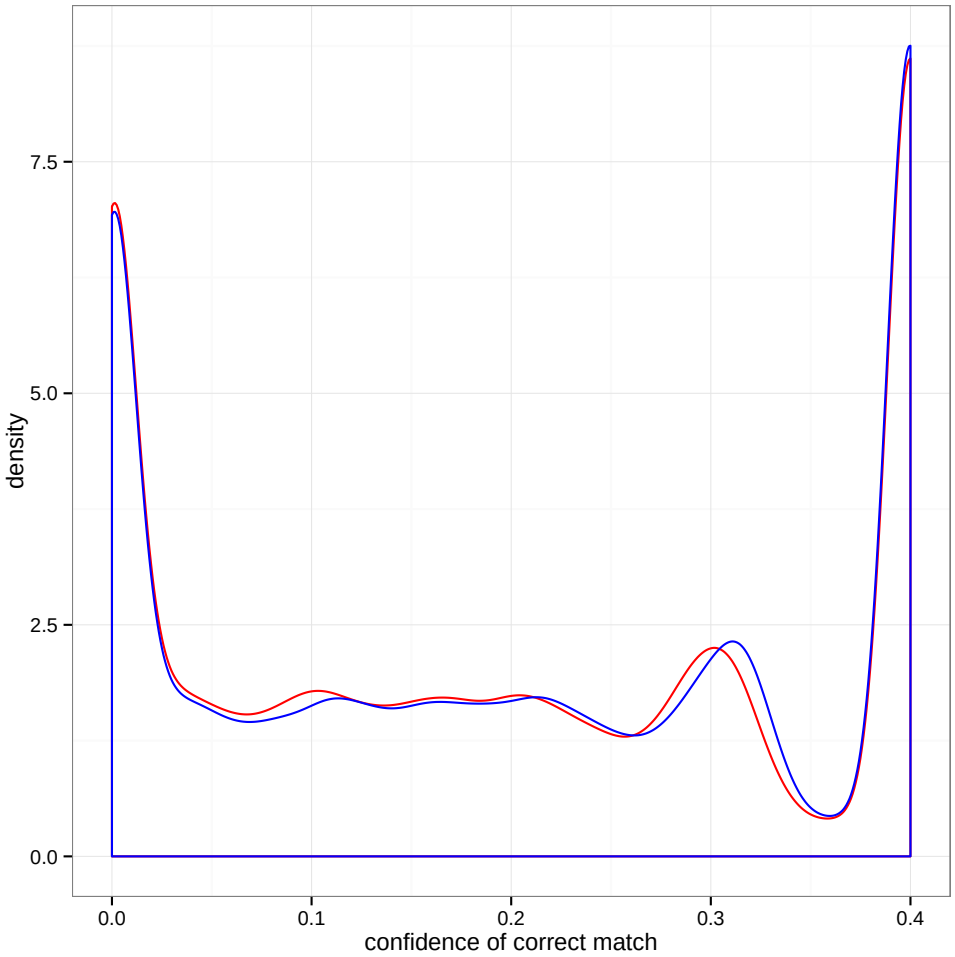


Figure 6.2: Density of confidence in correct matches in the All DAIs reference set for the Name-only approach (blue) and the Extended context approach (red).

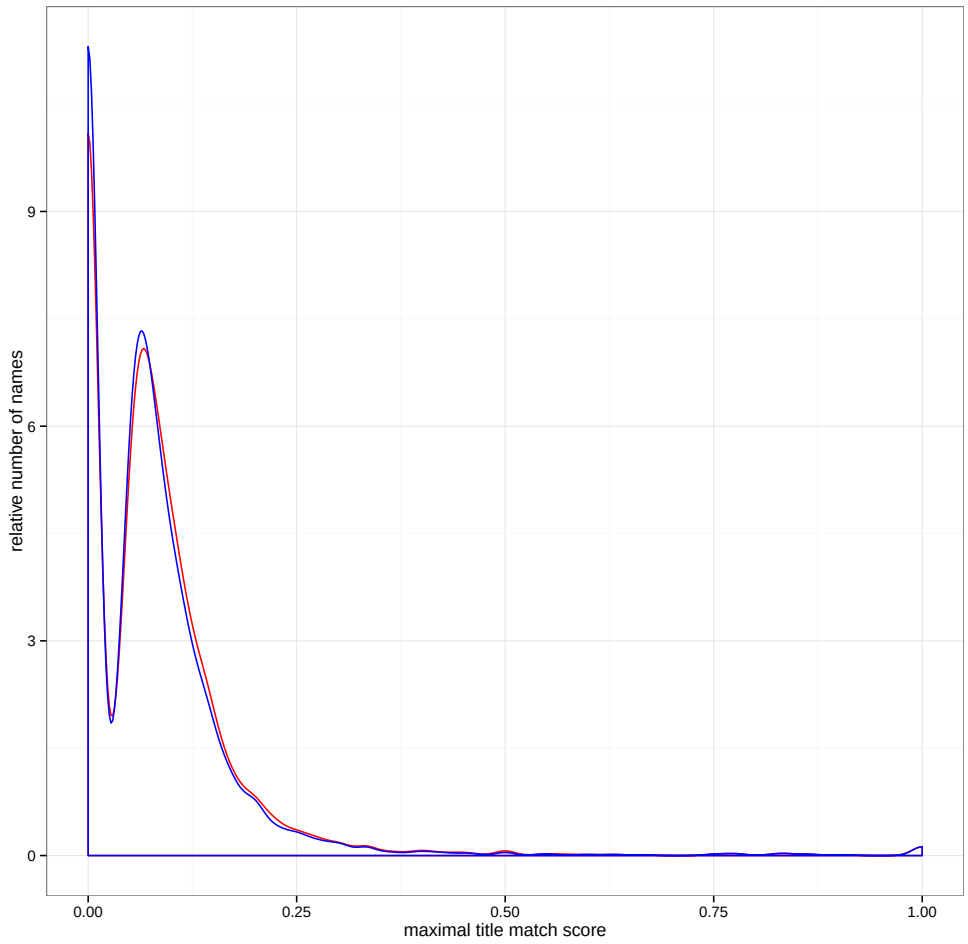


Figure 6.3: Density of title match score for the EEMCS reference set, for Dutch project titles (blue) and English project titles (red)

6.3 Conclusions

Setting the probability that no match is found to at least 0.6 puts the non-match always at rank 1. In the evaluation this yields only false negatives for the positive reference sets and only true negatives for the negative reference set. If we count only the match with the highest probability, the precision and recall of the positive and negative reference sets are on the extremes 0 and 1 respectively, making them hard to compare and draw a conclusion about the general precision and recall.

When we look at the second best match (ranked by probability), the precision and recall tell more about the results of the matching approach. About 78% of the matches in the positive reference sets (recall at rank 2 is 78.7% in the EEMCS set and 78.2% in the All DAIs set) are in this position, which is (precision at rank 2) 96.9% of the suggested not non-matches for the EEMCS set and 92.7% for the All DAIs set. The mean rank of the correct answer, counting missing matches as "rank infinite", is 2.5. 18.7% of the names (EEMCS set) or 15.6% (All DAIs set) did not match at all, suggesting that similarity threshold should be lowered, or that the blocking method was too strict. A quick glance at names that did not match, shows names in which the position of the surname prefix in the author name is different than *surname, surname prefix, given name* as used in the matching approach and author names of which the given name record is a combination of initials and full first name, next to spelling differences and single versus double family names.

On the other hand, only 34.4% of the names in the negative reference set did not match any researcher at all. The majority of the names in the negatives set is close enough to a name in the VSOI database that one or more matches are suggested. This suggests that matching should be more strict.

The Extended context approach did not change the results significantly. The observed differences can be ascribed to the random order of matches with equal confidences.

7 Conclusions

This final chapter summarises the research and results, answers the research questions and provides suggestions for future work.

To produce a complete overview of Dutch institutional science, DANS provides the NARCIS portal of publications from institutional repositories in the Netherlands (in the Index) and profiles of researchers, organisations and research projects (in the VSOI system). The publications' author names are not always linked to the researchers, because author names are usually ambiguous, and therefore identifiers are needed to unambiguously link to the correct researcher. However, the publication records do not include these Digital Author Identifiers for every name.
There are no links from publications to organisations or research projects that were involved in the production, because no identifiers are available for organisations and the grant agreement number as identifier for projects is of limited use and rarely used.

We developed two approaches to automatically match author names from publications in the Index to researchers in VSOI: the Name-only approach, which used a combination of three string comparison functions and normalisation to probabilistic matches; and the Extended context approach, which added comparison of publication titles and project titles as context to the name match.

In an experiment, the approaches were tested on a subset of the information available in the Index, namely records of publications from 2005 to 2012. The results of the experiment were evaluated using metrics that express the how many matches made are correct, based on reference sets of matches known from names with Digital Author Identifiers.

7.1 Conclusions

1. What do the data in the Index and VSOI database look like? What (potential) problems with data quality exist?

The Index consists of publication records in different metadata formats. Most records include a title, a list of author names, publication information, type of publication and other metadata. Of the available records, only those in MODS format can contain structured author names and provide DAIs along with the author names.

VSOI contains records about researchers, research projects and organisations. Many of these records are interlinked so that [the user can see who is involved in certain projects and what organisation employs a researcher or project manager]. Researchers who have a DAI can have that DAI added to VSOI so that the web portal can link publications with that DAI to the researcher's profile, which is based on the information in the researcher record.

Observed problems in the Index include (multiple) unstructured names in fields that should be structured, missing names and corporate names or non-names in personal family name fields. Not all DAIs found in publications are [real DAIs]: some are malformed, others are from different domains.

In VSOI, information is edited and checked manually. The snapshot of the database used in this research contained 47918 records about researchers. Errors are made, but it is impossible to tell how many. Because of the manual nature of updating and editing the database, it can take some time before errors are corrected. A problem that affected the research was the registration of five DAIs to two different researcher records each. Because it was hard to automatically identify the correct registration, and leaving the duplicate registrations in would mean identifying researchers by their DAIs was not always possible, the DAIs had to be removed from the reference sets.

The subset of records used for the experiment consisted of 276,394 publication records, containing a total of 1,010,848 names. In the experiment only the names with a family name field were input, resulting in an input of 930,647 names. 265,884 names were connected to a researcher via a DAI in the publication record. Among them were 9,851 names from researchers in the EEMCS faculty of the University of Twente, for which the DAIs are certainly correct in the publication records - in the larger set there may have been a few incorrectly registered DAIs. These set and subset were used as positive reference sets for the match results. Another 98,027 names had DAIs that were not registered in VSOI; these were used as a negative reference set.

2. How good is the result when just names are used?

The Name-only approach produced 5,611,824 matches with confidence > 0, for 585,925 names. More than one-third of the names did not match, which

was expected as the majority of all researchers (including many PhD students and researchers at foreign institutes) are not in VSOI. A manual test of two samples of 100 names each suggested that 60% of the names did not have a matching researcher in VSOI. This percentage was used in the algorithm as the probability that no researcher matches.

When the set of matches for a particular name is considered an answer and only the match with the highest probability (rank 1) counts, then all names are matched to the null researcher, i.e. all names are said to not match. That means none of the names in the positive reference sets are correctly matched (precision and recall are 0), but all of the names in the negative reference set are (precision and recall are 1).

When we look at the matches at ranks 2 and up, we see more interesting and meaningful results. Over 92% of the matches for names that should match a researcher at rank 2 is correct, and these are about 78% of all correct matches. About 82% of all correct matches were found and nearly all of these ranked in the top 6 by confidence.

3. What context is available? To what degree can an extended context approach improve a name only approach?

The Extended context approach did not create new matches, but tried to improve the results of the Name-only approach by differentiating the match scores using information from the context of the author names and researchers. Intuitively this is how humans would match names and researchers: can we be more confident in saying a researcher is the person belonging to this name, given the publication metadata and researcher's context (research projects and/or organisations)?

The context of author names is the publication metadata: title, publication type, title of publication, co-authors, etc. The context of researchers consists of assigned classifications, expertise and connections to research projects and organisations and their descriptions (title of research, name of organisation, classifications, co-workers and other information).

Based on observations that connections to research projects are the most widely available aspects of researchers (85.9% of the researchers have a connection to at least one research project), and that publications are often results of research projects on the same topic, we used a comparison of publication and project title as context extension. However, because the absence of a (partial) title match does not entail that there is no match, only the similarity's positive influence was used.

Unfortunately the Extended context approach did not improve the results. It could have improved the overall precision and recall scores when confidences of correct matches increased and confidences of incorrect scores decreased, but the differences are small enough that the random ordering of matches with the same confidences can explain why less correct matches are in ranks 2 to 6 for the EEMCS reference set and more correct matches are in ranks 2 to 6 for the All DAIs reference set.

94.4% project titles share less than a quarter of the words with the publication title. The matches that got a higher confidence because of matching title words were the incorrect matches. The confidences and hence results for matches with names from the negatives set did not change at all.

4. How do probabilistic matching results compare to non-probabilistic results?

The normal precision and recall show that only considering the top ranked match does not say anything about the matching performance. The second ranked match is the best result of the actual matching, but although the normal precision and recall show good results, not all matches at rank 2 are correct and certainly not all correct matches are at rank 2.

The probabilistic model provides a way for relating alternative possible matches and ranking them (except for equal probabilities), yet limiting the results to one answer per author name. The use of probabilities allows the outcome of the matching approaches to be used in other probabilistic calculations, such as classification of publications based on correlations between words in a publication title and classifications assigned to researchers.

We think that probabilistic matching is useful, but within the approaches described in this research, the probabilities (and when results are ordered by probability, ranks) are not assigned well. The expected precision and recall are clearly capped by the probability normalisation step that assigns 0.4 probability for a perfect match and 0.6 non-match for every name is not realistic, as the precision and recall for best matches show. Because non-match is the best match for every name, the expected precision and recall for the positive reference sets are only about 0.2. The expected precision and recall of 0.738 for the negative reference set is largely the result of the default 0.6 for non-matches. We expect that using more context elements can make the confidences more precise and realistic.

The E_{100} recall results show that about 82% of the correct matches in the All DAIs reference set were found and that most of these are among the top 6 matches for a name.

How can publications in NARCIS automatically be connected to their authors, organisations and research projects?

Publications can be automatically connected to at least some of the author researchers by matching author identifiers that were recorded in the publication record and in VSOI. Because most author names do not include identifiers, and it would take much coordination and hence a long time and manual work to solve this (if possible), automatic matching is more feasible.

The Name-only approach presented in this thesis matches author names to researcher names. The best matching researcher is in many cases correct. The Extended context approach with context from research projects needs to be designed differently. Publication titles and project titles do not match enough to differntiate correct and incorrect matches; adding more context (such as co-authors, co-workers or classifications of researchers, projects and organisations) could help.

In a public portal like NARCIS, completeness and correctness are important aspects. To increase completeness, DANS could add the best positive match as an option to the publication record and ask the matched researcher whether the publication is hers or his in a semi-automatic fashion.

Publication and researchers can connect to other entities in the research context via the researcher's existing connections. Research projects are not stored in VSOI forever and organisations may change (in which case the old records are not stored), so it will be harder to match to the correct entities. Using a probabilistic approach has the same (dis)advanteges as it has now: possibly large number of alternatives, but also smaller probability that correct matches are not found.

7.2 Recommendations

Recommendation 1 *Use different heuristics to determine probabilities of matching or non-matching.*

It appears that using a fixed prior probability of 0.60 that a name does not match at all does not work well in the probability that a name match is correct, resulting in a low expected precision and expected recall and 0 and 1 for normal precision and recall when looking at the best match (none of the correct positive matches matched best and all the correct negative matches matched best).

Recommendation 2 *Use tools to split names into functional parts, so that they can be sorted better.*

Unstructured names have been ignored in this research, because they were hard to split into functional parts (family name, given names, prefix, titles and affiliations). Automatically recognising these parts in surnames could improve the results of the name matching approach. Also, having a greater number of structured names would increase the applicability of the approach.

Recommendation 3 *Include dependencies in name matching.*

We have looked at name matching for each name separately. However, it stands to reason that certain combinations of matches are more likely than others. The first step could be to look at names related to potential matches, but more formal relations should be used. The probability that x_1 is the researcher behind name n_1 and the probability that researcher x_2 is behind name n_2 are expected to be interdependent. The researchers may know each other because they have co-authored before, work in the same research group or building, or may have participated in the same research project.

Recommendation 4 *Deduplicate publication records.*

Deduplication was discussed, but not applied in this research. If a set of records can be identified as semantic duplicates of one another and the information of these records integrated, some authors unidentified in one record can be identified by DAIs from the other record. This reduces the number of unidentified names and could help matching performance when using co-authors.

Recommendation 5 *Stimulate change in repositories; accept more types of identifiers.*

There are many people with an identifier or more than one identifier of different types. The DAI is only used in the Netherlands, whereas others like ORCID are gaining ground. An agreement on supporting identifiers other than DAIs should be made and implemented. Also, not all institutional repositories currently register DAIs for non-faculty staff, even though they allow publications by these employees in the repository and . This yields unlinked author names. Allowing DAI registration to all staff lets aggregators like NARCIS automatically create context for their publications.

Recommendation 6 *Repository software should encourage using conventions.*

MODS is mostly a structure definition for metadata records. Rules for interpretation of the record contents are made in separate standards or documents. This has lead to repository users or administrators, for example, to enter placeholders for people such as "et al.", for unknown publishers and place names such as "s.n." and "s.l.", which are treated as family names, publisher names and place names. Special values or, preferably, attributes to indicate missing information should be standardised and implemented in repository and OAI-PMH software, so that these can be used to handle the special meanings.

The MODS editorial committee has decided that future versions of MODS will contain an element to specify "et al." in a name [27], which is a first step.

Bibliography

[1] MPEG21 DIDL application profile for institutional repositories - standards - SURFwiki, April 2009. URL: http://wiki.surf.nl/display/standards/MPEG21+DIDL+Application+Profile+for+Institutional+Repositories.

[2] NARCIS classificatie. Technical report, KNAW, The Hague, September 2010. URL: http://www.narcis.nl/content/pdf/classification_nl.pdf.

[3] Science, February 2013. Page Version ID: 539201457. URL: http://en.wikipedia.org/w/index.php?title=Science&oldid=539201457.

[4] D. Barbará, H. Garcia-Molina, and D. Porter. The management of probabilistic data. *IEEE Transactions on Knowledge and Data Engineering*, 4(5):487 –502, October 1992. doi:10.1109/69.166990.

[5] Jeroen Bekaert, Patrick Hochstenbach, and Herbert Van de Sompel. Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory digital library. *D-Lib Magazine*, 9(11), November 2003. URL: http://www.dlib.org/dlib/november03/bekaert/11bekaert.html, doi:10.1045/november2003-bekaert.

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008. URL: http://iopscience.iop.org/1742-5468/2008/10/P10008, doi:10.1088/1742-5468/2008/10/P10008.

[7] P. Christen. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1537–1555, 2012. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5887335.

[8] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, page 73–78, August 2003. URL: http://dc-pubs.dbs.uni-leipzig.de/files/Cohen2003Acomparisonofstringdistance.pdf.

[9] Ben Companjen. Exploring NARCIS. 2012.

[10] DANS. About DANS, 2012. URL: http://www.dans.knaw.nl/en/content/about-dans.

[11] Elly Dijk, Chris Baars, Arjan Hogenaar, and Marga van Meel. NARCIS: the gateway to Dutch scientific information. In Bob Martens and Milena Dobreva, editors, *ELPUB2006. Digital Spectrum: Integrating Technology and Culture - Proceedings of the 10th International Conference on Electronic Publishing*, pages 49–58, Bansko, Bulgaria, 2006. URL: http://depot.knaw.nl/5631/1/Paper_ELPUB_2006.pdf.

[12] Uwe Draisbach and Felix Naumann. DuDe: the duplicate detection toolkit. In *Proceedings of the International Workshop on Quality in Databases 2010*, Singapore, 2010. ACM. URL: http://www.hpi.uni-potsdam.de/fileadmin/hpi/FG_Naumann/publications/2010/DuDe_-_The_Duplicate_Detection_Toolkit_cr.pdf.

[13] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1 –16, January 2007. doi:10.1109/TKDE.2007.250581.

[14] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049, doi:10.1080/01621459.1969.10501049.

[15] Anderson A. Ferreira, Marcos André Gonçalves, and Alberto H.F. Laender. A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record*, 41(2):15–26, 2012. URL: http://dl.acm.org/citation.cfm?id=2350040.

[16] Thomas Gurney, Edwin Horlings, and Peter van den Besselaar. Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, 91(2):435–449, 2012. URL: http://www.springerlink.com/content/3267244176v663x3/abstract/, doi:10.1007/s11192-011-0589-1.

[17] Mauricio A. Hernández and Salvatore J. Stolfo. The merge/purge problem for large databases. *ACM SIGMOD Record*, 24(2):127–138, May 1995. URL: http://doi.acm.org/10.1145/568271.223807, doi:10.1145/568271.223807.

[18] Arjan Hogenaar and Wilko Steinhoff. Towards a dutch academic information domain. In *Third International Conference on Open Repositories 2008*, Southampton, United Kingdom, 2008. URL: http://pubs.or08.ecs.soton.ac.uk/12/.

[19] In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. On co-authorship for author disambiguation. *Information Processing & Management*, 45(1):84–97, January 2009. URL: http://www.sciencedirect.com/science/article/pii/S0306457308000721, doi:10.1016/j.ipm.2008.06.006.

[20] Jasper Kuperus. *Catching Criminals by Chance: A Probabilistic Approach to Named Entity Recognition using Targeted Feedback*. Master's thesis, University of Twente, Enschede, the Netherlands, 2012. URL: https://mail-attachment.googleusercontent.com/attachment/?ui=2&ik=a4e9ba9ccc&view=att&th=13818955a1a74ef8&attid=0.1&disp=safe&zw&saduie=AG9B_P_uvnFAAcoZKoBMjWiRh48E&sadet=1340718142928&sads=XX-5LmNtsnBWgnAtV5ISc-rDrPA.

[21] Hanna Köpcke and Erhard Rahm. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2):197–210, February 2010. URL: http://www.sciencedirect.com/science/article/pii/S0169023X09001451, doi:10.1016/j.datak.2009.10.003.

[22] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. Open archives initiative - protocol for metadata harvesting - v.2.0, June 2002. URL: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

[23] Dongwon Lee, Jaewoo Kang, Prasenjit Mitra, C. Lee Giles, and Byung-Won On. Are your citations clean? *Communications of the ACM*, 50(12):33–38, December 2007. URL: http://doi.acm.org/10.1145/1323688.1323690, doi:10.1145/1323688.1323690.

[24] Library of Congress, Network Development and MARC Standards Office. Metadata object description schema: MODS, May 2013. URL: http://www.loc.gov/standards/mods/.

[25] Matteo Magnani and Danilo Montesi. A survey on uncertainty management in data integration. *Journal of Data and Information Quality*, 2(1), July 2010. doi:10.1145/1805286.1805291.

[26] Paolo Manghi and Marko Mikulicic. PACE: a general-purpose tool for authority control. In Elena García-Barriocanal, Zeynel Cebeci, Mehmet C. Okur, and Aydın Öztürk, editors, *Metadata and Semantic Research*, volume 240 of *Communications in Computer and Information Science*, pages 80–92. Springer Berlin Heidelberg, 2011. URL: http://www.springerlink.com/content/u628643038530q67/abstract/.

[27] MODS Editorial Committee. Changes for MODS version 3.5, February 2013. URL: http://www.loc.gov/standards/mods/changes-3-5.html.

[28] F. Panse, M. van Keulen, and N. Ritter. Indeterministic handling of uncertain decisions in deduplication. *Journal of Data and Information Quality*, not available yet, 2012. URL: http://eprints.eemcs.utwente.nl/21610/.

[29] Jeroen Salman, Maarten Kleinhans, Dolf Weijers, Gerben Bekker, Mirjam Perry, and Marion de Boo. *Kennis over publiceren.* De Jonge Akademie, Amsterdam, December 2012. URL: http://www.knaw.nl/Content/Internet_KNAW/publicaties/pdf/20121015.pdf.

[30] Dorothea Salo. Name authority control in institutional repositories. *Cataloging & Classification Quarterly*, 47(3-4):249–261, 2009. URL: http://www.tandfonline.com/doi/abs/10.1080/01639370902737232, doi:10.1080/01639370902737232.

[31] Sumit Sarkar and Debabrata Dey. Relational models and algebra for uncertain data. In Charu C. Aggarwal, editor, *Managing and Mining Uncertain Data*, pages 45–76. Springer, 2009.

[32] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles. Efficient topic-based unsupervised name disambiguation. In *International Conference on Digital Libraries: Proceedings of the 2007 conference on Digital libraries*, volume 18, page 342–351, 2007. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.3737&rep=rep1&type=pdf.

[33] Li Tang and John Walsh. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3):763–784, 2010. URL: http://www.springerlink.com/content/e016u5q71jh36540/abstract/, doi:10.1007/s11192-010-0196-6.

[34] Annemiek van der Kuil and Martin Feijen. The dawning of the Dutch network of Digital Academic REpositories (DARE): a shared experience. *Ariadne*, (41), October 2004. URL: http://www.ariadne.ac.uk/issue41/vanderkuil.

[35] Maurice van Keulen and Ander de Keijzer. Qualitative effects of knowledge rules and user feedback in probabilistic data integration. *The VLDB Journal*, 18(5):1191–1217, 2009. URL: http://www.springerlink.com/content/v1t230411x852680/abstract/, doi:10.1007/s00778-009-0156-z.

[36] Maurice Vanderfeesten, Magchiel Bijsterbosch, and Lisanne Boersma. info-eu-repo - standards, 2011. URL: http://wiki.surf.nl/display/standards/info-eu-repo.

[37] Ellen M. Voorhees. The TREC-8 question answering track report. In *Proceedings of TREC*, volume 8, pages 77–82. NIST, 1999. URL: http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf.

[38] William E. Winkler. Overview of record linkage and current research directions. Technical Report #2006-2, U.S. Census Bureau, Washington, DC, USA, 2006.