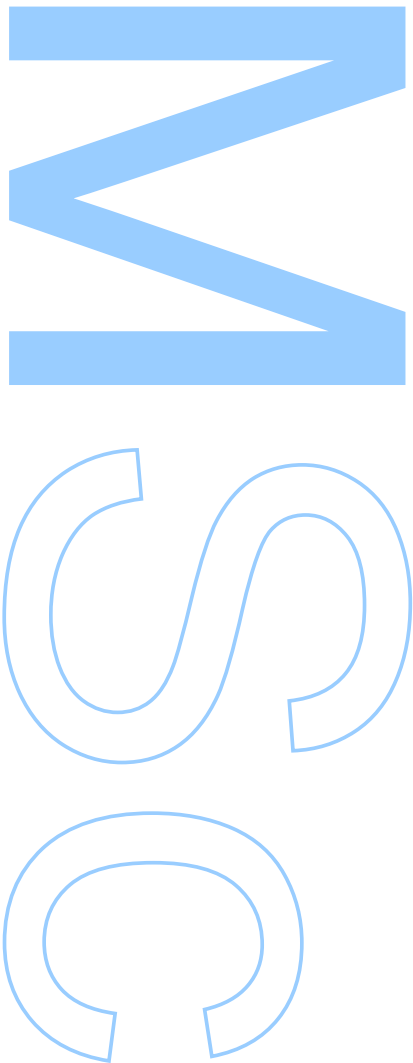
Automatic identification of institutions in affiliation strings

José Pedro Ribeiro Azenha Rocha
Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos
Departamento de Ciência de Computadores
2020

Orientador
Professor Álvaro Pedro de Barros Borges Reis Figueira,
Faculdade de Ciências da Universidade do Porto



U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

UNIVERSIDADE DO PORTO

MASTERS THESIS

Automatic identification of institutions in affiliation strings

Author:
José Pedro Ribeiro Azenha ROCHA

Supervisor:
Álvaro Pedro de Barros Borges Reis FIGUEIRA

A thesis submitted in fulfilment of the requirements
for the degree of MSc. Network and Information Systems Engineering

at the

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

December 15, 2020

Acknowledgements

I would like to thank Professor Álvaro Figueira, for his knowledge, availability to answer questions and for allowing me to learn from this new and educative experience.

I would also like to thank Sylwia Bugla from the Authenticus project for always being available to answer questions about Authenticus and helping me understand the platform.

I am thankful to INESC TEC and Fundação para a Ciência e a Tecnologia for the scholarship provided in order to do this work (project UIDB/50014/2020).

Finally, I would like to thank my parents, for always providing me with everything I needed and for supporting me through my educative journey.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

MSc. Network and Information Systems Engineering

Automatic identification of institutions in affiliation strings

by José Pedro Ribeiro Azenha ROCHA

Universities and research centers try to improve their scientific performance in order to get a better position in institutions' rankings and to be able to receive more funding. For the measuring of scientific output of these institutions, it is fundamental that normalized and validated data is provided by bibliographic databases, mainly the identification of institutions associated with publications. The affiliation of the author contained in a publication's metadata describes which institution the author is connected to by the means of an unformated piece of text. Multiple ways of describing an institution exist, and each bibliographic database uses its own format and various styles which may change over time.

From this context arose a need to create an algorithm which would automatically identify a real word institution from an author's affiliation string. This dissertation presents such a novel algorithm, allowing scientific publications to be automatically associated to their corresponding institutions, and thus count towards those institutions' performance statistics and bibliometrics. The development of the algorithm resulted in an API software which takes as an input an affiliation string and it successful outputs institutions identified from it.

In order to build the algorithm, an analysis of the available data and affiliation strings was made. From these analysis, three different methods were created. The choice of a method to use is done using regular expressions which determines if certain components exist in the string. If an email address is present, than the email based method is used for the identification. In case the string is or contains an ISNI number, the corresponding

method is used. When none of the two cases applies, a general method, which uses n-grams and tf-idf in order to learn from a training dataset and predict the correct institution for new affiliation strings, is used.

Due to the different characteristics of each of the three methods, different evaluation methodologies were used. We evaluated how many email addresses were able to directly or indirectly identify an institution. For the ISNI identification component, we tested how many identifications were correct, incorrect or how many failed. The last n-grams based method of the algorithm was evaluated using a train-test split and cross-validation of the train dataset.

The results show that our methods are, generally, very effective in identifying institutions from affiliation strings. The email based identification successfully identified 87% of cases. The ISNI based identification method correctly identified 78%, with only 3% of false positive cases. The n-gram based identification method showed very good results when identifying higher education institutions (F1-score of 0.95), while the identification of research centers had more modest results.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos

Identificação automática de instituições em frases de afiliação

por José Pedro Ribeiro Azenha ROCHA

Universidades e centros de investigação tentam melhorar o seu desempenho científico para conseguir obter uma melhor posição em rankings de instituições e para conseguirem mais financiamento. Para a medição de produção científica, é fundamental que sejam fornecidos dados normalizados e validados por parte das bases de dados bibliográficas, principalmente a identificação de instituições associadas a publicações. A afiliação de autor contida nos meta-dados de uma instituição descreve qual a instituição a qual o autor está relacionado, pelo meio de um fragmento de texto não formatado. Contudo, existem múltiplas maneiras de descrever uma instituição e cada base de dados bibliográfica usa o seu próprio formato e vários estilos que podem mudar ao longo do tempo.

Daí surge uma necessidade para criar um algoritmo que identifica automaticamente uma instituição real a partir da frase de afiliação de um autor. Esta dissertação apresenta um novo algoritmo, permitindo associar automaticamente publicações científicas às instituições correspondentes, e contar para as estatísticas bibliométricas e de desempenho dessas instituições. O desenvolvimento do algoritmo resultou num software API que recebe como input uma frase de afiliação e retorna com sucesso instituições identificadas na frase.

Para construir o algoritmo, uma análise dos dados disponíveis e das frases de afiliação foi feita. A partir desta análise três métodos diferentes foram criados. A escolha de que método usar é feita usando expressões regulares que determinam se certo componente existe na frase. Se existir um endereço de email, o método baseado no email é usado para a identificação. Caso a frase seja ou contenha um número ISNI, o método correspondente é usado. Quando nenhum dos dois casos se aplica, um método geral, que usa n-grams

e tf-idf para aprender a partir de um conjunto de dados de treino e prevê a instituição correta para uma nova frase de afiliação, é usado.

Devido às diferentes características de cada um dos três métodos, diferentes metodologias de avaliação foram usadas. Avaliamos quantos endereços de email conseguiram identificar diretamente ou indiretamente uma instituição. Para a componente de identificação usando ISNI, testamos quantas identificações foram corretas, incorretas ou não foram possíveis. O último método baseado em n-grams foi avaliado usando uma divisão treino-teste e cross-validation do conjunto de dados de treino.

Os resultados mostram que os nossos métodos são, geralmente, muito eficazes a identificar instituições a partir de frases de afiliação. A identificação à base de email identificou com sucesso 87% dos casos. O método de identificação à base de ISNI identificou corretamente 78% dos casos, apenas com 3% de falsos casos positivos. O método de identificação à base de n-grams mostrou resultados muito bons ao identificar instituições de ensino superior (F1-score de 0.95), tendo alguns problemas com a identificação de centros de investigação.

Contents

List of Tables

List of Figures

Acronyms

APA American Psychological Association.

API Application programing interface.

CERIF Common European Research Information Format.

CRACS Centro de Sistemas de Computação Avançada.

CRIS Current Research Information System.

DBLP Digital Bibliography & Library Project.

DGEEC Direção-Geral de Estatísticas da Educação e Ciência.

DGES Direcção-Geral do Ensino Superior.

DOI Digital Object Identifier.

FCT Fundação para a Ciência e a Tecnologia.

HTTP Hypertext Transfer Protocol.

INESC TEC Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência.

ISBN International Standard Book Number.

ISNI International Standard Name Identifier.

ISO International Organization for Standardization.

ISSN International Standard Serial Number.

JSON JavaScript Object Notation.

LSI Latent semantic analysis.

MARC Machine-Readable Cataloging.

MODS Metadata Object Description Schema.

NLP Natural language processing.

NUTS Nomenclature of Territorial Units for Statistics.

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting.

ORCID Open Researcher and Contributor ID.

REST Representational state transfer.

RIS Research Information Systems.

SHA Secure Hash Algorithm.

SQL Structured Query Language.

TF-IDF Term frequency-inverse document frequency.

URL Uniform Resource Locator.

WOS Web of Science.

XML Extensible Markup Language.

Chapter 1

Introduction

Universities and research centers struggle every day to improve their scientific performance and appear on a high position on research institution's rankings and reports. On the other hand, funding government bodies also require trustful sources of information about research output of institutions in order to steer research management. To enable credible and reliable bibliometric studies and analysis of the scholar outputs it is fundamental that bibliographic databases of scientific publications provide normalized and validated data. From the institutional point of view, the most important validation is the identification of real world institution associated with publications. In a paper[1], an overview is done of how publication metrics are used to evaluate impact and scholarly productivity by institutions and researchers and how these measures can impact future funding allocation. While the paper describes both traditional and new techniques to measure impact and productivity, all of these use publication data as a base, whether it is citation count, online views, or another metric.

1.1 The Authenticus project

Authenticus is a software platform[1] developed at the University of Porto and CRACS/INESC TEC, that aims to build a national repository of publications metadata authored by researchers of Portuguese institutions. The system uploads publications from multiple indexing databases and automatically associates publication authors with

[1] https://www.authenticus.pt/

known researchers and institutions. It has been designed and implemented to provide researchers and institutions with a set of functionalities and specialized interfaces to manage their scientific data and confirm or dismiss associations proposed by the automatic methods. Authenticus allows interoperability with other systems, provides synchronization with ORCID, both for import and export, among many other functionalities. It currently contains 574k of publications, over 86k researchers and more than 2k of Portuguese and worldwide institutions.

One of the key features of the Authenticus system is the possibility of importing publications from multiple sources in an automatic way. Publications metadata are uploaded every week from Web-of-Knowledge, Scopus, DBLP and ORCID, processed by specialized algorithms and made available for researchers and institutions to validate. The increase of new publications in Authenticus is presented in figure 1.1 [1]. On average, there are around 800 new records imported every week.

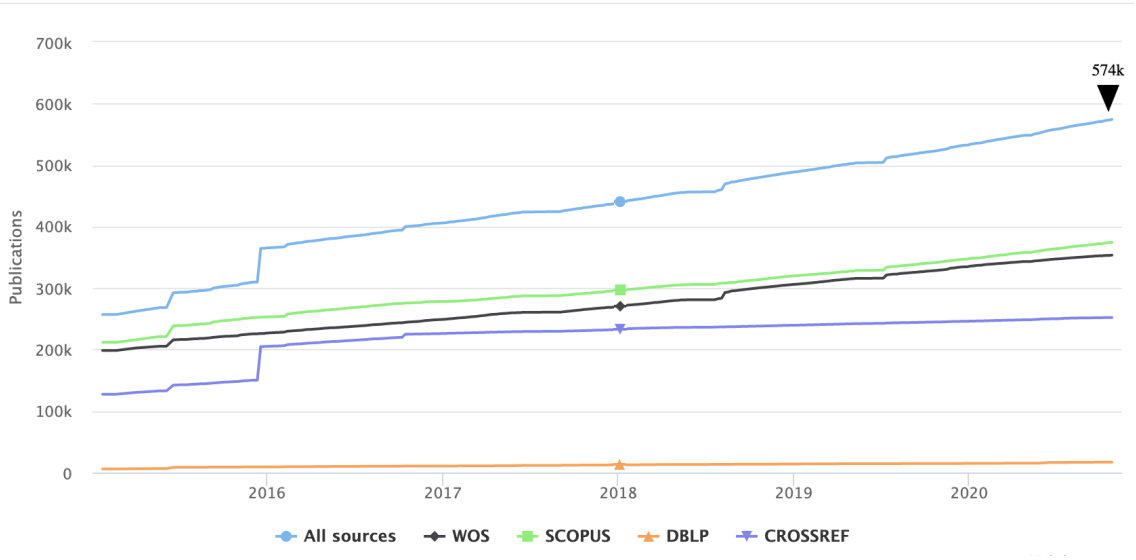


FIGURE 1.1: Increase of new publications in Authenticus

The metadata retrieved from external sources is stored in a database in a raw format and is a subject of various processing and analysis. This work analyses the Authenticus publications metadata, specifically the author's affiliation string in order to find institutions described in it.

[1] https://www.authenticus.pt/en/publications/statistics

1.2 Motivation

The metadata of a publication is a structured bibliographic information which provides details about a publication and helps to identify it. The metadata includes such a information as: author's names, publication title, publication type, affiliation of the author, year of publication among others. The amount of information depends on the format in which it is represented and on the bibliographic database. The element of the publication metadata which is the most appropriate for identifying institutions is the affiliation of the author.

The author's affiliation is usually a simple unformated piece of text, that describes the institution to which the author is connected, very often including the city and country of the institution. There is no homogeneous way of describing one institution. Each bibliographic databases has its own format and uses various abbreviations and styles which additionally may change over time.

In order to allow bibliometric studies focused on institutions, it is necessary to match each affiliation of an author with the real world institution and location. This is the only way to enable credible rankings of institutions on the national or world level. It allows to find out, for example, what is the most important institution in a specific field, what is the institution's research volume, income and reputation or, by measuring citations, to find out what is the institution's research influence.

In the specific case of Authenticus, the vast majority of the 574k existing publications are yet to be associated to institutions. Additionally there are around 800 new publication records being added to the database every week. These factors show that manual association of the publications to institutions is an incredible arduous task, if not impossible. Because of this, automatic association is needed.

1.3 Objectives

The main goal of this project is to develop an algorithm/API that automatically associates scientific publications with the respective universities and/or research centers, using affiliation strings extracted from publication metadata.

In this project we have intermediate several goals to achieve:

- Analyze the available publications metadata, specifically the affiliations strings, in order choose the strategy for building the algorithm.

- Create the algorithm of institution identification which detects the specific characteristics of the affiliation string and applies a appropriate identification method.
- Evaluate all the components of the algorithm for institution identification.
- Implement the novel solution in the Authenticus system infrastructure.
  The approach was to build an API (application programming interface).

The work focuses on Portuguese institutions and organizations but its scale can be broaden to a worldwide level.

1.4 Methodology

Firstly, a study of the background and state of the art of the subject was made, in order to understand what type of strategies should be used. It also helped understand where the metadata comes from and how the affiliation strings are usually formatted.

We were provided with tables from Authenticus database relating to institutions, publications and affiliation strings. These tables will be described in chapter 3 and later in section 5.1.1.

We observed the data available to us to understand what information we could use to help in identification. We also chose types of affiliation strings that could be better identified using one method or another. We settled with strings which are email addresses, strings which contain an ISNI number and other strings.

We started creating the methods for each type, using Python as the programming language, evaluated them and observed the initial results. After this, we made decisions on what could be improved in the methods. Finally, we evaluated the final methods with the objective of determining how good they are at identifying institutions successfully.

1.5 Novelties and Contributions

In this document we will describe our approach for identification of institutions in author affiliations.

This approach introduces several novelties, namely, identification and disambiguation of institutions using an email address and identification of institutions using n-grams in conjunction with tf-idf.

We also make several contributions by creating methods for identification using email addresses, ISNI numbers and n-grams in conjunction with tf-idf, a method for disambiguation of institutions with the same email address, a general algorithm that chooses the best method for identification and an API for using the algorithm.

1.6 Structure

This document is divided into six chapters. The first chapter gives a explanation of the problem and how we attempted to solve it. Chapter 2 discusses background and work that has been done relating to our problem. In chapter 3 we describe the datasets from the Authenticus database used in our work. Chapter 4 explain in detail our approach towards solving the problem. In Chapter 5 we explain how the evaluation of our work was done, present the final results, discuss the limitations of our work and describe how the API for the experimental case works. Finally, chapter 6 resumes the contributions made and lists some future work that could be done in order to improve our results.

Chapter 2

Background and Related Work

This chapter starts with an explanation of issues related with scientific publications, bibliographic metadata formats and types. We give an overview of the tools and APIs available which deals with the processing of publication's metadata. Later we discuss the known approaches in the literature for identification of institutions from affiliation string or other metadata.

In the last part we describe some natural language processing (NLP) techniques commonly used for named entities identification.

2.1 Scientific publications and bibliographic data

Scientific publications, or commonly known research papers, are written documents published in scientific journals which are the most important means of communicating results and knowledge derived by science. In recent decades, high-quality paper writing has become a key qualification of scientists and institutions. Publications are cataloged in libraries and electronic databases and are characterized using extended document metadata which carries all relevant information, such as document title, abstract, publication year, and authors. This metadata builds a bibliography record which consists of a bibliographic data described in bibliographic format.

In the last years academic publishing has transitioned almost entirely from the print to the electronic format. Bibliographic databases started to play an important role in cataloging and indexing the information, exposing the data to outside to and providing bibliometric measures, such as number of citations. Indexing databases started to appear more often and almost every scientific areas had its own. One of them are open with the

data fully available for the public, such as CrossRef[1], DBLP[2] or arXiv[3]; others are commercials and access to them is regulated with policies and payments (WOS[4] or Scopus[5]). Together with the databases started to appear various human and/or machine readable bibliography formats whose purpose is not only to store information but also to exchange in a form of references. Nowadays the know bibliographic formats are: BibTex[6], APA[7], RIS[8], EndNote[9], MODS[10]or Machine-Readable Cataloging (MARC)[11]. The formats differ in semantics, syntax and complexity of the data they provide.

2.1.1 Data Interoperability

Interoperability means "the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality"[2]. To facilitate data inseparability between different bibliographic databases, standards frameworks and data formats started to appear.

Some of the interoperability frameworks and formats include:

CERIF XML - is the exchange format inspired by the CERIF - The Common European Research Information Format. CERIF is the comprehensive information model for the domain of scientific research. It is intended to support interchange of research information between and with CRIS system. CRIS - current research information system "is a database or other information system to store, manage and exchange contextual metadata for the research activity funded by a research funder or conducted at a research-performing organization."[12]

OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting[13] is an application-independent protocol, allowing translating metadata into a common core set of elements and exposing them for harvesting. This protocol can be used by data

[1] https://www.crossref.org/
[2] https://dblp.uni-trier.de/
[3] https://arxiv.org/
[4] https://webofknowledge.com/
[5] https://www.scopus.com/
[6] http://www.bibtex.org/
[7] https://apastyle.apa.org/
[8] https://web.archive.org/web/20170707033254/http://www.researcherid.com/resources/html/help_upload.htm
[9] https://endnote.com/
[10] https://www.loc.gov/standards/mods/
[11] https://www.loc.gov/marc/
[12] https://en.wikipedia.org/wiki/Current_research_information_system
[13] https://www.openarchives.org/pmh/

providers, to expose structured metadata, and by service providers, for making requests to retrieve said metadata.

OpenURL - provides a standardized format for transporting bibliographic metadata about objects between information services. This format works similarly to a standard URL. However, instead of referring to a website, it refers to a resource within a website, such as an article or book. It is also supposed to work as a permalink, meaning that it remains the same for a given resource, regardless of the hosting website.[3]

2.1.2 Metadata exposing tools

Many of the indexing and aggregating bibliographic databases provide access to the data by specialised API (Application programing interface). The API allows other systems and applications to connect to the database registry, including reading from and writing to records. Some API systems are freely available to anyone, others are provided only with membership subscription. We list a description of the API for some of the most used bibliographic databases:

ORCID public API - publicly available API, with registration required, which returns ORCID IDs and data made public by ID holders. This API allows searching by bibliographic, affiliations, funding, research activities and ORCID record data, or all types at the same time. The API returns an XML file containing ORCID IDs of records holding public data that matches the search query.

CrossRef API - publicly available REST API which exposes the metadata deposited in CrossRef database. Besides bibliographic metadata it also provides access to metadata about funding data or journals. You can search, facet, filter, or sample metadata from thousands of members, and the results are returned in JSON format.

Scopus API - paid REST API provided by Elsevier that retrives data from Scopus, its citation database. In addiction to other standard information, it provides other useful information, such as citation data and abstract, full journals and books published by Elsevier and research metrics. It is able to search using affiliation strings, author name, DOI and other identifiers and by a generic search query.

WOS API - paid API provided by Clarivate with data from publications since the year 1900. While it has data also provided from free services, it also has useful paid information, such as journal impact factor, author variant names, grant information, abstract and others.

This services provide a great and easy way to access and reuse scientific publications metadata in a structured and organized way to be used in institutional repositories and research networking systems. Although they all follow the same architectural style (REST - Representational state transfer), they differ in syntax, semantics and complexity of data they provide. It is relatively easy to implement the external application interfaces in the existing system, but there is no specialized tool which is able to uniquely process the received metadata.

The data exposed by the services differs depending on the accessibility of the system. The free services provide less data and the paid one, more.

The common fields exposed by the free and paid services include:

- DOI
- Abstract
- Authors' names
- Author IDs (for example ORCID IDs)
- Authors' affiliations
- Author Keywords
- Document title
- Document publication year
- Source (journal) title
- ISSN
- ISBN
- Source type (journal, conference proceeding, etc.)
- Volume/issue/pages/article number

- Document type (e.g. research article, review article, etc)
- Publisher

Additionally the paid systems provide specialised information about:

- Source/Journal Metrics (Cite Score or Impact Factor)
- Subject categories
- Citation count (number of times cited by other articles)
- Cited works

2.2 Known approaches to institutions' identification

The specific problem of identifying institutions from publication's affiliation string has been already tackled by various groups of researchers. This section gives and overview of the techniques used to solve the problem of identifying institution entity.

2.2.1 Rule-based algorithms

The most basic approach to the problem of identifying entities relies on heuristics and handcrafted rules extracted from unstructured text. Rule-based algorithms receive as an input an unstructured text, which in a first phase passes through text processing phases. Text processing transforms text into something an algorithm can digest and usually involves processes such as tokenization, removing unnecessary punctuation, tagging, removing stop words, stemming or lemmatization. From the processed text the algorithm applies various rules and filters in order to discover which entity the input text corresponds to.

An example of rule-based algorithm to identify the institution is a method described in [4, p. 55], where every affiliation string in a publication, passes through two steps, preprocessing and identification. The identification process builds a SQL query, executes it on the database table of known Portuguese higher education institutions and research centers, and, at the end, evaluates the result of the query.

Scientific literature provides numbers of examples where rule-based named-entity recognition methods for knowledge extraction were applied. Development of highly accurate rules is extremely time consuming, prone to error and very domain specific. Once

built, a rule-based system does not perform well on other domains [5]. An advantage of such system is that it can be highly accurate, however constant supervision is required in order to adapt rules to text modification [6].

2.2.2 Web supported rule-based identification

Another interesting approach for institution identification was presented in a paper [7], describing a method that uses an approximate string metric that handles acronyms and abbreviations, by measuring similarity comparing the result sets of web searches.

This method measures the similarity between two strings, in this case, the affiliation string to be identified and the name of an institution that might be the corresponding one, by searching both in a web search engine (e.g. Google, Yahoo) and observing how many of URLs from the given results overlap between the two strings searched. This number of similar URLs was then normalized and formed the similarity score.

This approach proved to give better results than other similarity techniques like Levenshtein distance and trigram. Levenshtein distance is a metric for understanding the similarity between two strings. It works by calculating the minimum number of changes required to change one string into the other. These changes can be either the insertion, substitution or removal of a single character. The lower the score is, the more similar the strings are. Trigram is a case of n-grams where n is 3. N-grams will be explained in section 2.3.1.

This was an expected result, considering the fact that the other techniques calculate the similarity based on the text itself and not on its meaning and, therefore, the chance of success is lower, specially with cases like abbreviations and different organizations with similar names.

Another advantage of using a web search engine is resilience to typing errors which can occur in affiliation strings and which usually are well handled by search engines.

The main disadvantage of this method is its low accuracy for new institutions since these yield a lower amount of search results. It is also dependent on third party companies for it's accuracy, which means that sudden changes in the algorithm of search engines could impact significantly the performance of the method.

2.2.3 Standard Institution Identifiers

One of the best solutions for entity identification is the usage of standard identifiers. Nowadays, there are various commonly used identifiers for researchers and academics (ORCID or CiênciaID), for scientific works (DOI) or for example for journals (ISSN) . Few years ago there has been also introduced a standard open identifiers for organizations, namely ISNI[1] and Ringgold ID.

Open ISNI for Organizations

The International Standard Name Identifier (ISNI) is the globally recognized and adopted international standard approved by ISO (ISO 27729) for the unique identification of the public identities of persons and organizations across all fields of creative activity.

To understand the potential of the ISNI number and its usage for identifying institutions, it is necessary to look into its older 'sibling' - ORCID number[2]. ORCID number, a persistent digital identifier for researcher, was introduced in 2009 and 11 year latter it is widely used by research community for author disambiguation especially for connecting researchers with their works, awards, and affiliations.

Nowadays, ISNI ID is already used for sharing public information about identities online and in databases, across stakeholders, national borders and in the digital environment. The ISNI community is made up of several constituencies: the founding members, registration agencies, general members, data contributors, and organizations concerned with the identification and description of information resources (such as ORCID)[1].

An agency Ringgold[3] which provides a quality data to power scholarly communications is a first ISNI Registration Agency for organizations. Ringgold released a free service[4] to provide open access to the ISNI Identifiers and data for organizations. The service includes: (1) an API to obtain and resolve existing ISNIs for organizations, (2) a complete dataset download of ISNIs, organization names, locations, alternate names, and (3)URLs and a free online look-up service to search and obtain ISNI records. The agency states that the identify database includes over 500,000 ISNI numbers for organizations, representing 99.9% coverage.[3]

ISNI is a relatively fresh invention, however it is already used by publishers and bibliographic indexing databases. With the ongoing growth in scholarly publishing, the usage

[1] http://www.isni.org/content/isni-community
[2] https://orcid.org/
[3] https://www.ringgold.com/isni/
[4] https://isni.ringgold.com/

of the organizational identifier allows to track publications across whole institution or within a school or department, can reduce confusion resulting from name changes, mergers and name variants including names translated into other languages. It seems to just be a matter of time when ISNI number will be attached to a metadata describing scholarly outputs and from there accurate and reliable identification of institutions is only a step further.

Our algorithm for identifying institution in an affiliation string of a scientific publication, is enriched with a functionality of ISNI look-up. The input of this method is an ISNI number, which first we consult locally in Authenticus database and later in the Ringgold database using the API the agency provides.

Ringgold ID

The Ringgold ID[1] is a proprietary identification standard, with the goal of identification and disambiguation of organizations in the scholar field, created by the Ringgold agency. It consists of four to six digits, currently, and is issued in numerical order, meaning that the identification number corresponds to the order in which the identification record was created.

2.3 NLP Techniques for Institution Identification

Institution identification could be considered a subset of Natural Language Processing, which studies ways to extract information from natural text. While the first is a very specific problem with little research, the latter is a very active field for scientific research. For this reason we decided to look at some of the techniques used in this broader field to assess if they could be applied in our specific context.

2.3.1 N-grams

An n-gram is a contiguous sequence of n words from a given text. Typically, they are formed by splitting a string and forming the grams. First, the string is split into the different words. Then, the n contiguous words starting with the first one make the first n-grams. The second n-gram is composed of the words starting at the second word and so on. A visual representation of the n-gram splitting, using 3 as the value for n (trigrams), of the string ”This is a simple piece of text” can be seen in figure 2.1.

[1] https://www.ringgold.com/ringgold-identifier/

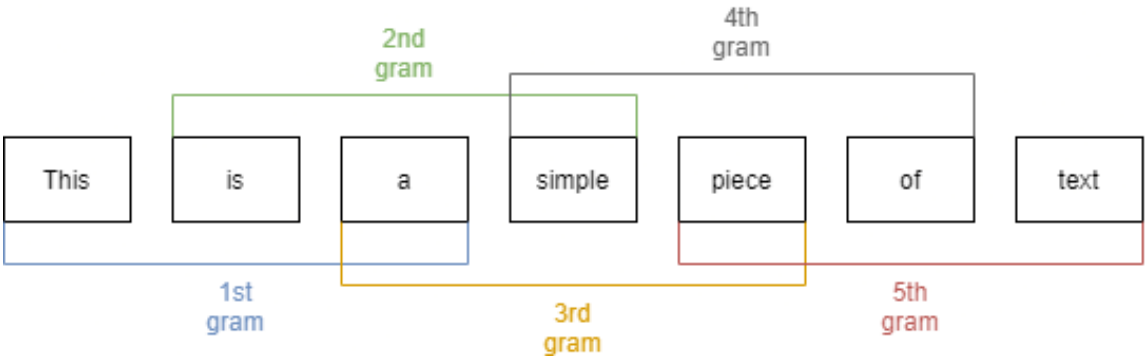


FIGURE 2.1: Trigram splitting of a string

N-grams can be used for statistical natural language processing, by using their frequency for classifying text, for example. By using n-gram frequency instead of single word frequency, more context of the class can be gathered. While several classes can have the same words, the bigger the n of n-grams, the less documents have the same grams. If n is too small, the grams will not be specific enough and wrong matches will happen (underfit). If n is too big, the grams will be too specific and other examples of the same class will not be able to be identified (overfit). In [8], a method for text categorization is described and shown to have very good results and little computational power required. This method uses samples for the desired categories, builds a profile for each by ordering the generated n-grams by highest frequency. The paper also refers to some shortcomings of the method, mainly that the performance of the matching is greatly impacted by the quality of the training set and that for the method to be effective, a good form of normalization of the results is required. In [9], n-grams are used at character level for authorship attribution of anonymous text with language independence, achieving state of the art performance. In the majority of the results presented in this paper, the optimal n ranges between 3 and 5.

2.3.2 TF-IDF

TF-IDF, or "Term frequency-inverse document frequency", is a numerical statistic that tries to show how important or relevant a word is in a document in a collection of documents [10]. It works by taking into account both how common a word is in a given document and how rare it is across all documents of the collection and can be explained by the following equation:

$$tfidf(t,d,D,N) = tf(t,d) * \log \frac{N}{D}$$

where

- t: term
- d: document
- D: number of documents that contain t
- N: total number of documents
- tf(t,d): raw frequency of t in d

This value increases proportionally to the frequency of a term in a document and is offset by the number of different documents that contain the term.

In a paper [11] a comparison is made between tf-idf, lsi (latent semantic analysis) and multi-words, showing that, despite having less semantical quality, tf-idf has more statistical quality for text classification.

In [12], an experiment is made with the task to automatically generate summaries from text. Using extraction technique (use relevant words from the text to generate the summary), tf-idf is shown to be a good method to create a value that shows how important a word is in a document.

2.4 Cross Validation

Cross-validation is a technique used to evaluate how the results of a statistical analysis for a given model will generalize for an independent dataset. The way the technique works is by the means of splitting the given dataset into k folds (parts) of equal size and, for good results, equal representation of the classes present in the dataset as a whole. With these folds, an evaluation of the model is made using one of the folds as the testing data and the remaining ones as the training data. This process is repeated until all folds were used as the testing data. Finally, an average of the scores is made and the final result is given.

In [13] it is shown that cross-validation, while being computationally more intensive (the algorithm has to run for each one of the k folds), provides a better assessment of fit of a model for new data than regular train-test split.

2.5 Metrics

Precision

Precision is a metric that represents, from predictions for a given class, how many are true. This metric can be calculated by dividing the number of true positives (cases where the prediction of the class was true) by the number of true positives plus the number of false positives (cases where the prediction of the class was false).

Recall

Recall is a metric that represents, from all the cases where the actual result was the class, how many predictions where true. It can be calculated by dividing the number of true positives by the number of true positives plus the number of false negatives (cases where the actual result was the class, but another class was predicted).

F-Score

General f-score is a performance metric generated from the harmonic mean between precision and recall and can be shown by the following equation:

$$\mathrm{F} = (1 + \beta^2) \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{(\beta^2 \cdot \mathrm{Precision}) + \mathrm{Recall}}$$

where β is the number of times recall is considered more important than precision.

F1-score is the most common version of the metric, where the value for β is 1, meaning recall and precision have the same importance. It can be shown by the following equation:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Confusion Matrix

The previous metrics are calculated when there is a classification that is either positive or negative. For situations where there can be several resulting classes (multi-class classification), a confusion matrix is needed to calculate the metrics for the dataset as a whole.

Confusion matrix is a table layout where the rows are the predicted classes and the columns are the actual classes, or vice-versa, in which the values represent the number of examples with a specific actual class and a specific predicted class. A simple visual representation can be seen in figure 2.2, where we have classes 1, 2 and 3 and n(i,j) indicates the number of examples of i that where predicted as j.

Actual \ Predicted	1	2	3
1	n(1,1)	n(1,2)	n(1,3)
2	n(2,1)	n(2,2)	n(2,3)
3	n(3,1)	n(3,2)	n(3,3)

FIGURE 2.2: Confusion table representation

For calculating the final metrics, an average of each metric for each class is done. This average can be macro, simply summing the results of each class and dividing by the number of classes, weighted, multiplying the results by the number of cases for each class and diving the whole by the number of total cases, and micro, by calculating the metrics with the sum of cases for all classes (true positives, false positives, true negatives and false negatives).

Chapter 3

Datasets Structure

This chapter describes in detail the datasets used to build and evaluate our institution identification algorithm. The algorithm seeks to identify association of author's affiliations string extracted from metadata of a publication to corresponding real world institutions or organizations. First we describe the main sources of data, its formats and structures, how the data is pre-processed and give examples of the common strings of affiliations that are being identified. The last section of this chapter focuses on the description of the related datasets required to build the algorithm, namely a dataset of institutions and a dataset of publications.

3.1 Data Sources

Our algorithm for institution identification has been optimized for author's affiliations strings exported from external databases. The import of the data is handled by the Authenticus system, which pre-processes and stores the metadata in separate tables.

Authenticus, as a publications metadata aggregator, can currently import publications metadata from three external sources: Web of Science (WOS), Scopus, CrossRef and two other external aggregators: ORCID and DBLP. Each of the metadata providers uses different APIs, thus the information retrieved from the external systems has a different structure, different data characterization and classification.

The process of importing metadata from various sources is presented in figure 3.1. It involves two steps:

Step 1 Get publications metadata from external source using different systems APIs;

Step 2 Store raw publication metadata: a group of actions that process the retrieved metadata and stores the raw data in the Authenticus database for further processing;

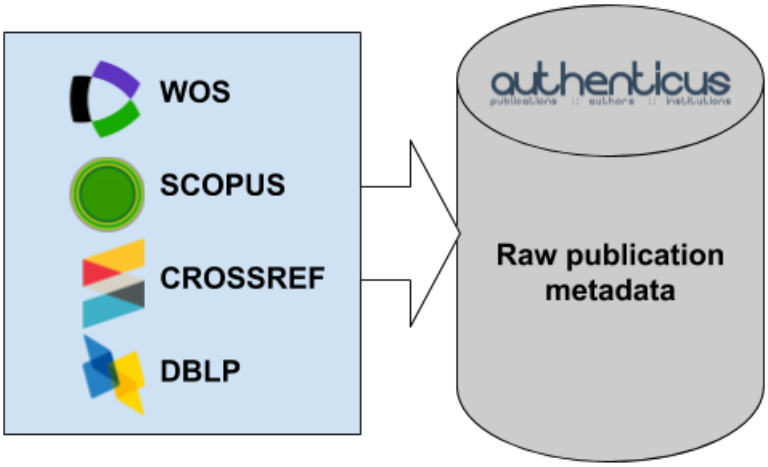


FIGURE 3.1: Representation of the metadata import process

Raw publication metadata is later a subject of the pre-processing method, which extracts different elements from the original metadata, normalizes it, and stores in various tables. In the next sub-sections we describe the flow of the affiliations string from the original format, through pre-processing until the final format used by the algorithm. In figure 3.2 we can see a diagram of the datasets used for the algorithm.

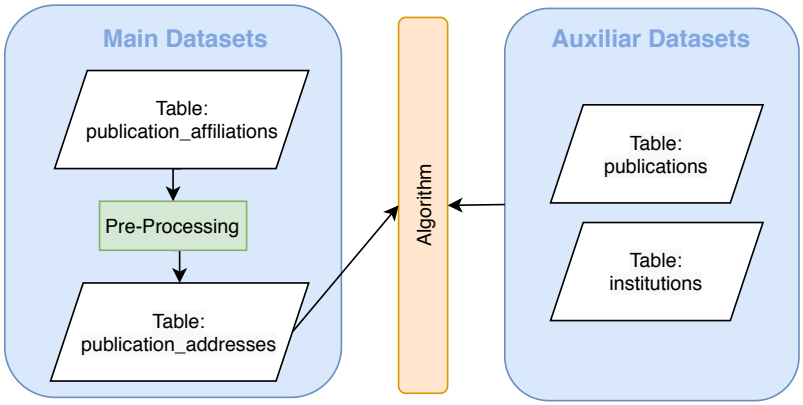


FIGURE 3.2: Dataflow diagram of the datasets

3.1.1 Dataset of affiliations

Affiliation strings extracted from publication metadata are stored in a raw format in a table called `publication_affiliations`. The structure for this table is presented in table 3.1.

One publication may have various affiliation strings associated to it. Each string of affiliations has one source type which indicates the origin of that metadata. The possible source types are: scopus, wos, cross-ref, dblp, wos_reprint and scopus_reprint. The last column in the table, *preprocessed*, indicates whether the affiliation string is already pre-processed or not. The pre-processing process is described in section 3.1.2.

Name	**Description**	Total Entries: 9715006 Entries per type: • scopus: 7331440 • wos: 1697370 • scopus_reprint: 312774 • wos_reprint: 290445 • crossref; 82967 • dblp: 10
id	Id for the entry	
publication_id	Id of the publication in table `publications` (section 3.3.2)	
affiliation	Affiliation string	
type	Source from where the affiliation string comes from ('scopus', 'wos', 'crossref', 'dblp', 'wos_reprint' or 'scopus_reprint')	
preprocessed	Boolean flag for whether the string has been pre-processed and is available in the `publication_addresses` table	

TABLE 3.1: Data structure for table `publication_affiliations`

Tables 3.2 shows examples of publication metadata extracted from two different sources: WOS and Scopus. The important parts related to affiliation strings are in bold. These strings of affiliations are stored in a separate table: `publication_affiliations` which is a base for further processing. Table 3.3 presents how the data from the source metadata in the above example is processed and stored in the `publication_affiliations` table. In the next step the data from that table is an object of pre-processing method which is described in the following section.

Metadata from WOS

{ "publication_type": "J",
"authors": "Collaco, P; Silva, JCE",
"authors_full_name": "Collaco, P; Silva, JCE",
"document_title": "A complete comparison of 25 contraction conditions",
"publication_name": "NONLINEAR ANALYSIS-THEORY METHODS & APPLICATIONS",
"document_type": "Article; Proceedings Paper",
"author_address": "Univ Aveiro, Dept Math, P-3800 Aveiro, Portugal; Univ Coimbra, Dept Math, P-3000 Coimbra, Portugal",
"reprint_address": "Collaco, P (reprint author), Univ Aveiro, Dept Math, P-3800 Aveiro, Portugal.",
"e_mail_address": null,
"year_published": "1997",
"digital_object_identifier": "10.1016/S0362-546X(97)00353-2"}

Metadata from SCOPUS

{ "Authors": "Collaco P., Carvalho E Silva J.",
"Title": "A complete comparison of 25 contraction conditions",
"Year": "1997",
"Source title": "Nonlinear Analysis, Theory, Methods and Applications",
"Affiliations": "Department of Mathematics, University of Aveiro, 3800 Aveiro, Portugal; Department of Mathematics, University of Coimbra, 3000 Coimbra, Portugal",
"Authors with affiliations": "Collaço, P., Department of Mathematics, University of Aveiro, 3800 Aveiro, Portugal; Carvalho E Silva, J., Department of Mathematics, University of Coimbra, 3000 Coimbra, Portugal",
"Correspondence Address": "Collaço, P.; Department of Mathematics, University of Aveiro, 3800 Aveiro, Portugal",
"Abbreviated Source Title": "Nonlinear Anal Theory Methods Appl",
"Document Type": "Review",
"Source": "Scopus" }

TABLE 3.2: Example of metadata imported from WOS and Scopus for the same publication

Univ Aveiro, Dept Math, P-3800 Aveiro, Portugal	wos
Univ Coimbra, Dept Math, P-3000 Coimbra, Portugal	wos
Collaço, P., Department of Mathematics, University of Aveiro, 3800 Aveiro, Portugal	scopus
Carvalho E Silva, J., Department of Mathematics, University of Coimbra, 3000 Coimbra, Portugal	scopus

TABLE 3.3: Example of affiliation strings after being extracted from metadata sources

3.1.2 Data Pre-processing

The data pre-processing module processes, parses and standardizes the affiliations string in order to extract important, contextual information that can be useful for the identification algorithm to classify the affiliation.

The process receives as an input the raw affiliation string and the source/type of the string (WOS, Scopus, etc). For each type of affiliation string the pre-processing varies, due to the different formats the string has at its origin. The affiliation string may contain an institution abbreviation, country and city. Very often it also has indications to which author it belongs. The information that the pre-processing algorithm may extract from the strings are:

- Author names
- City
- NUTS II region[1]
- Country
- Abbreviations (abbreviations of research units or schools, for example)

The methods to extract the contextual information from the affiliations string is usually based on a string matching algorithm which tries to match various tokens of the affiliation string with real entities, for example countries. In many cases the elements of the affiliation string follows an order, where at the beginning appears names of the authors associated, then the institution string, city and country. This and other observations allowed us to create rules which help to extract detailed information from the affiliation's strings. Examples of the string can be find in the section 3.2.

[1] https://en.wikipedia.org/wiki/NUTS_statistical_regions_of_Portugal#NUTS_II

The pre-processing module stores the data in a new table `publication_addresses`, which is described in the next sub-section.

3.1.3 Dataset of pre-processed affiliations

The data extracted in the pre-processing phase is stored in a table called `publication_addresses`. The structure of this table is presented in table 3.4.

Name	Description
id	Id for the entry
publication_id	Id of the publication in table `publications` (section 3.3.2)
institution_string	Part of the affiliation string identified as being part of the institution name
institution_string_sha1_hash	SHA-1 hash for the institution_string
author_string	Part of the affiliation string identified as being part of the author's name
country_id	Id of the country that was identified in the string
region_id	Id of a NUTS II region that was identified in the string
city	Name of a city identified in the string
abbreviation	Abbreviation for an institution identified in the string
type	Source from where the affiliation string comes from ('wos', 'scopus', 'dblp', 'email', 'crossref', 'scopus_reprint', 'wos_reprint', 'email_reprint')

TABLE 3.4: Data structure for table `publication_addresses`

The pre-processed data stored in this table is an input for the algorithm to identify affiliations. The most important fields from this table are `publication_id`, `institution_string`, `country_id` and `city`, the first one being used for gathering additional data about the publication and the others directly for identification.

3.2 Affiliation Strings

An affiliation string is a type of metadata of a publication that describes the affiliation between an author and the institution the author was working at the time of writing the

publication. The affiliation string is a simple string of text without a standard identifier for the institution, but very often written with some rules and in a specific order. The rules are provided by bibliographic databases such as WOS, Scopus, CrossRef or DBLP and each of the string source have its own format.

String Examples

There is no single, nor obvious, way to describe the affiliations string. One institution can be described with abbreviations, in a full formal, mixed (some parts abbreviated some in full format), and in different languages. Many of the strings follow a specific order of elements which characterise the institution, such as authors names, followed by institution name, city and finally country, for example.

Affiliations can be presented with one of three main types:

- address text with the author's name first
- address text without the author's name
- email address

One string can have one of the types or a combination of them, usually having one email address maximum. Examples of these strings can be observed in table 3.5.

Type	**Examples**
Address text with author name	Machado Miguens, J., Laboratorio de Instrumentacao e Fisica Experimental de Particulas-LIP, Lisboa, Portugal;
	McDonald, J.; CFMC-GTAE, Av. Prof. Gama Pinto 2, Lisboa 1699, Portugal;
	Pinto De Carvalho, A., Cad. Urol., Fac. Med., Lisboa, Portugal.
Address text without author name	FAC MED LISBON,DOENCAS PULMONARES CLIN,LISBON,PORTUGAL;
	Clinical Pharmacology Unit, Instituto de Medicina Molecular, Lisbon, Portugal2Laboratory of Clinical Pharmacology and Therapeutics, Faculty of Medicine, University of Lisbon, Lisbon, Portugal5Center for Evidence-Based Medicine, Faculty of Medicine, Univer;
	Univ Pinhal Marrocos, P-3030290 Coimbra, Portugal.
Institutional Email Address	aoteles@fc.up.pt
Address text with multiple affiliations	David, L., Inst. of Molec. Pathol. and Immunol., University of Porto, IPATIMUP, 4200 Porto, Portugal, Inst. of Molec. Pathol. and Immunol., University of Porto, IPATIMUP, Rua Dr. Roberto Frias s/n, 4200 Porto, Portugal;
	[Jimenez-Valverde, Alberto] Univ Azores, Azorean Biodivers Grp, Angra Do Heroismo, Portugal;
	UNIV PORTO, CTR EXPTL MORPHOL, P-4000 OPORTO, PORTUGAL.

TABLE 3.5: Examples of the affiliation strings in the Authenticus databases.

As mentioned previously in this section, there are multiple ways an institution can be described. In table 3.6 some examples of strings that have been identified as belonging to University of Porto can be seen. In examples 1 and 2 of the table, either the portuguese word for the city (”PORTO”) or the international word for the city (”OPORTO”) is used to refer to the university. In example 5 the word ”CIENCIA” is used as part of the description of the Faculty of Sciences, while in example 6 only an abbreviation of the english equivalent (”SCI”) is used. In examples 1, 2, 3 and 5 the word for the city after the faculty is used to refer to the university, while an explicit designation to it is used in examples 4, 6 and 7.

Number	Example
1	FAC MED OPORTO, INST HISTOL & EMBRYOL, P-4200 PORTO, PORTUGAL
2	SOARESDASILVA, P (reprint author), FAC MED PORTO,FARMACOL LAB,P-4200 PORTO,PORTUGAL.
3	FAC ENGN PORTO,CTR ENGN QUIM,RUA BRAGAS,P-4099 PORTO,PORTUGAL
4	FREITAS, V (reprint author), UNIV OPORTO,FAC ENGN,GABINETE CONSTRUCOES CIVIS,OPORTO,PORTUGAL.
5	CORREA, CMMD (reprint author), FAC CIENCIAS PORTO,CTR INVEST QUIM,PORTO,PORTUGAL.
6	LAGE, EJS (reprint author), UNIV PORTO,FAC SCI,DEPT PHYS,P-4000 PORTO,PORTUGAL.
7	UNIV PORTO,PORTO,PORTUGAL

TABLE 3.6: Examples of affiliation strings for the University of Porto

3.3 Related Database Resources

To build the affiliations matching algorithm it was necessary to use additional resources which provides more precise information about the strings being identified or about the source/publication to which the string was attached.

This section previews the complementary resources of the database that are used by the matching algorithm and provides description on how the data of these sets is used.

The complementary resources are represented by the following tables:

- Institutions
- Publications

3.3.1 Institutions

The institutions table includes over 11000 records, which represents over 800 Portuguese institutions (and their sub-units) of higher education and research centers from the private and public sector. It also has over 9000 universities from around the world, which were not covered in the scope of this work.

This dataset makes part of the Authenticus database. It was initially created with publicly available resources, such as data of DGES (Direcção-Geral do Ensino Superior) or DGEEC (Direção-Geral de Estatísticas da Educação e Ciência), and in case of the world universities data from Github[1]. During the years of functioning of the Authenticus system, the dataset was populated with new institutions wherever it was necessary.

The information about institutions is organized hierarchically, meaning that we can see if institutions have some sub-units or if it is already a parent institution. The dataset includes several tables, where the main one is a table of `institutions` and some auxiliary tables to represent the hierarchical structure, different types or areas, and overtime names structure modifications. Table 3.7 shows the structure of the `institutions` table, with only the fields revelant for the algorithm.

Name	**Description**
id	Id of the institution
name	Name of the institution
english_name	English version of the name of the institution
URL	Official URL domain of the institution
email_domain	Official email domain of the institution
isni_id	Official ISNI number of the institution
postal_code	Postal code of the institution's location

TABLE 3.7: Simplified structure for table `institutions`

The algorithm is using the data of this table in order to extract the name, email domain and start and end years of activity of institutions relevant to a prediction.

3.3.2 Publications

The publications dataset is the main set of the Authenticus database. The dataset contains over 540 000 publication metadata records. This table is populated by importing data from external resources, such as WOS, Scopus, CrossRef or DBLP. The process of importing the data into Authenticus is executed weakly and it brings around 700 new records every week.

The structure of the table is complex, as it has over 50 columns, holds the publication primary key used all over the Authenticus application and includes details such as title,

[1] https://github.com/endSly/world-universities-csv

year published, pages, volume number etc. The entire publications ecosystem has over 30 tables with complementary data, connected each other with weak or strong constrains.

The algorithm connects with this table in order to extract the year of a publication, in order to compare with the start and end years of activity of the institutions that might be in question for the prediction.

Chapter 4

Algorithm

In order to identify institution from metadata affiliation strings we propose an algorithm whose goal is to deliver the best possible guess from a set of institutions in a database. In this chapter we discuss in detail the structure and implementation of the identification algorithm. Basically, our method receives as input an affiliation string in the form of a tuple with an institution string, a city name and a country name and outputs the identified institution. Along this chapter we will refer to ”institution string” when mentioning the specific string itself and to ”affiliation string” when mentioning the tuple as a whole.

This chapter is divided in four parts. The first section describes the general idea of the algorithm, focusing on a top level processes and data flows. The following parts describe various components of the main algorithm, starting with the email based method and then, goes the ISNI method. The last part provides details of the *n-grams* algorithm implementation. It provides descriptions of the methods and solutions required to apply the *n-grams* algorithm including the string pre-processing, grams calculation and prediction step. Each step is additionally supported by a flowchart diagram.

4.1 Top Level Algorithm

The goal of the algorithm is to provide an automated method which successfully and uniquely identifies real world institution from author’s affiliation string provided in a publications metadata.

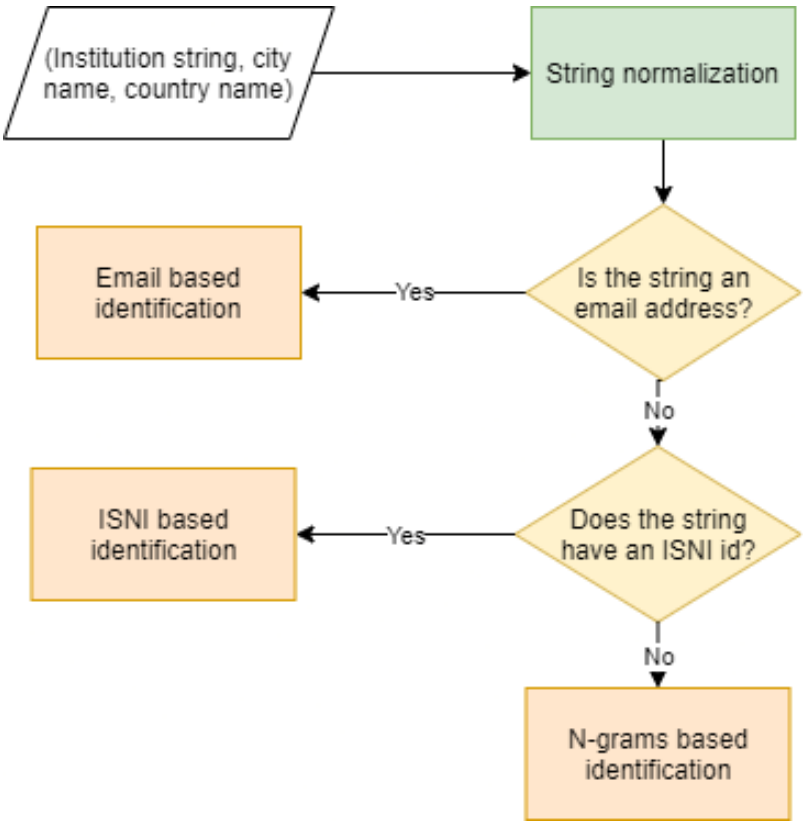


FIGURE 4.1: Diagram of the top level of the algorithm

The diagram of the top level of the algorithm is presented in the figure 4.1.

Input data

The algorithm receives as an input a set of strings that describes the institution. The set is extracted from the `publications_address` database table and it may consist of elements such as: institution string, city name and country name. The city and country name elements are extracted in the initial pre-processing described in the section 3.1.2. It some cases, the affiliation string is not very detailed and the city and/or country name cannot be extracted. In that situation the initial set consist only of one element - the institution string.

String normalization

The initial stage of the general algorithm pre-processes the strings in order to normalize their format. The normalization processes includes transformation of the strings into a well defined and consistent form to be used in the later identification process. The process includes changing the case to lowercase and convert special alphabetic characters into regular characters. The rest of the string transformation was already done

in the previous pre-processing method (sec 3.1.2) which ends in storing the data into `publications_address` database table.

Further processing of the string is made later in the n-gram identification method, which is described in section 4.4.

Method selection

The institution string, after being normalized, is used as a base to choose the best method for the identification of institutions. The selection method uses regular expressions and determines which of the identification methods to use. In order to better understand how these regular expression work, an explanation of the symbols used in regular expression can be seen in table 4.1.

Symbol	Meaning
.	Matches any single character except the end of a line.
^	Means the match starts at the beginning of the text.
$	Means the match ends at the end of the text.
[ ]	Matches any character inside the brackets.
-	Is used to represent a range of letters or number, typically used inside square brackets.
( )	Groups regular expression.
{ }	If they contain a number, matches the exact number of times the preceding character. If they contain two numbers seperated by a comma, matches the preceding character if it repeats a number of times between the two number. If they contain a number and a comma, the preceding character repeats at least the number of times.
|	Matches either the regular expression preceding it, or the regular expression following it.
?	Matches whether the character preceding it appears.
*	Matches the character preceding it 0 or more times.
+	Matches the character preceding it 1 or more times.
!	Does not match the character or regular expression following it.
\	Uses the next character literally.

TABLE 4.1: Regular expressions symbols

The three methods are:

E-mail based identification - The first regular expression check verifies weather the institution string is an email address. The regular expression used here is the following:

```
^[a-z0-9_.+-]+@[a-z0-9-]+\.[a-z0-9-.]+$
```

If the verification is successful, then email based identification is chosen. Full description of the email identification method is provided in the section 4.2. In case the email verification fails, then the algorithm proceeds to the second check.

ISNI based identification - The second check determines whether the institution string contains an ISNI identification number that identifies public identities of individuals and organizations. The specification of the ISNI ID is described later in this chapter (section 4.3). The regular expression used in this verification is the following:

```
([0-9]{15}[0-9X])|([0-9]{4} [0-9]{4} [0-9]{4} [0-9]{3}[0-9x])
```

Successful match of the ISNI in the institution string decides about the selection of the second identification method: ISNI based identification which is described in detail in the section 4.3. If the ISNI verification fails, the n-gram based identification is used.

N-gram based identification - This identification method is selected when both checks, email and ISNI verification are unsuccessful. The n-gram based identification uses n-grams in conjunction with tf-idf frequency information of the grams per institution in order to identify new affiliation strings. The full description of the method is provided in the section 4.4

Final output

The final output of the algorithm to identify the real-world institution for the affiliation string, for both the email and the ISNI method is (1) an identified institution in a structure of a sequence or (2) null value in case no institution was identified. The sequence structure represents the hierarchy of the institution from the lower, identified sub-institution until the top level, parent institution. In case the top level institution is identified, the output sequence has only one element. In the opposite case, meaning the identified institution is a sub-institution, then the output sequence contains all elements from the sub-institution until the top level element. Three examples of this can be seen in table 4.2.

Input	Output
none@fe.up.pt	[Faculty of Engineering of University of Porto, University of Porto]
none@up.pt	[University of Porto]
none@yahoo.com	null

TABLE 4.2: Example of outputs with sequence structure

In the case of the n-grams identification method, the output is a pair composed of one top level high education institution and one top level research center, where each one can be null or have an institution identification number. Three examples of this can be seen in table 4.3, with the institution identification numbers replaced by the names of the institutions.

Input	Output
Univ Lisbon, Inst Mol Med, Lisbon, Portugal	['University of Lisbon','Institute of Molecular Medicine']
UNIV COIMBRA,DEPT QUIM,P-3409 COIMBRA,PORTUGAL	['University of Coimbra',null]
CTR FIS NUCL,P-1699 LISBOA,PORTUGAL	[null,'Center of Nuclear Physics']

TABLE 4.3: Example of outputs for the n-gram method

The reason for this limitation will be described in section 5.1.1.

Technical details

Our algorithm implementation was written in Python 3.8 and uses several libraries, such as:

- re - Python's default regular expression library. This library is used throughout the algorithm for regular expression pattern detection and string transformation.
- json - Python's default library for JSON format data handling. We use this library to convert data to and from JSON format for easier data storage in text based files.
- pandas - Fast and efficient open-source library for data analysis and manipulation. This is one of the main tools of the algorithm. It is used to load data from datasets and manipulate data.

- nltk - The Natural Language Toolkit provides several libraries for text processing and interfaces for corpora and lexical resources. It provides the method for generating n-grams and a list of Portuguese and English stop-words.
- requests - HTTP requests library for Python. Allows to request data from third party APIs.

The Authenticus database has the tables mentioned in an SQL server, but for this work these where exported into csv files, which are then loaded and manipulated using the pandas library.

In the remainder of this chapter we present details of each of the identification methods and brief descriptions of the data and tools used.

4.2 E-mail Based Identification

The method for the email based identification is simple, but if successful, it can identify an institution without resorting to the other more complex procedures. The idea is to verify whether the domain of the email address can identify an institution.

Email addresses are composed by three parts: the first one is a string that identifies the person in the domain, the second is the symbol "@" that serves as a separator and the third is the domain. The domain of the email is a host name to which the email message will be sent. It is very common, especially in case of the emails extracted from the metadata of a scientific publication, that the domain of the email can identify an institution. The Authenticus database keeps track of email domains for institutions in the `institutions` table (section 3.3) and that gives us a base to search for domains and identify the correct institution.

Table 4.4 shows three examples of email addresses whose domains can be used to identify the institution.

Email address	**Host Institution**
090402101**@fep.up.pt**	Faculty of Economics of the University of Porto
aao**@esmae.ipp.pt**	School of Music and the Performing Arts of the Polytechnic Institute of Porto.
apa**@deq.uc.pt**	Department of Chemical Engineering of the Faculty of Science and Technology of the University of Coimbra.

TABLE 4.4: Examples of email address whose domains identifies the institution.

A diagram for the email based identification is presented in figure 4.2. The algorithm is only triggered if the institution string that is being identified is an email address.

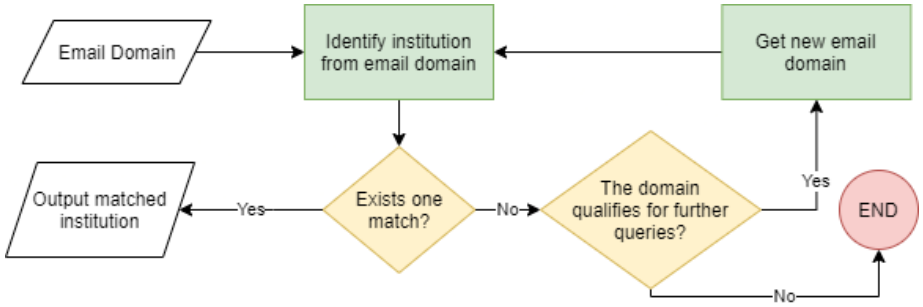


FIGURE 4.2: Diagram of the email based identification method

When an email address is identified, the algorithm extracts its domain and inputs into the email identification method. The method starts by searching in the database if the given domain exists. The table `Institutions` contain data about the host name stored in a column `email_domain`. We use this data to try to match the institution, by using the following pandas transformation:

```
institutions[institutions.email_domain == email_domain]
```

where "institutions" is the dataframe that contains the `institutions` table and "email_domain" is the email domain used for the query.

Using the email domain also has the advantage of being able to understand the hierarchy of institutions and get levels of identification, due to it's nature. These domains have a structure such that each element (string separated by dots ".") belongs to the element at it's right. If no institution is found with the initial email domain, we try and keep removing the left most element from it until either we can identify an institution or the domain ends up with less than two elements.

One email domain can have more than one institution related to it, since some institutions use the same email domain for sub-units as well. If this situation happens, we try to compare the email domain with the URL of the institution and see if they are strictly identical. For this we use the regular expression "`^https?:\/\/(www\.)?email_domain\/?$`", where "email_domain" is the email domain of the institution. With this regular expression we retrieve the institions where their URL matches the regular expression using the pandas transformation "`institutions[(institutions.url.str.match(regex) == True)]`", where "institutions

is the dataframe which contains the `institutions` table and "regex" is the regular expression. If this step does not identify a singular institution (returns none or more than one institution), we apply the same procedure as if no institution was found and try to identify a single institution higher up in the hierarchy.

In table 4.5 we can see an example of the recursive procedure for the non-existent email domain "faketext.dcc.fc.up.pt" and what would be the institution identified in each of the iteration.

Match iteration	**Email domain**	**Matched Institution**
1	faketext.dcc.fc.up.pt	-
2	dcc.fc.up.pt	Department of Computer Science of the Faculty of Science of the University of Porto
3	fc.up.pt	Faculty of Science of the University of Porto
4	up.pt	University of Porto

TABLE 4.5: Example of iterations of an email domain search for "faketext.dcc.fc.up.pt"

If an institution is identified, we use the same method as described before of recursively searching for a higher domain in order to generate the sequence structure.

The output of the email identification method is an institution in a sequence structure or null if no institution is identified.

4.3 ISNI Based Identification

ISNI - International Standard Name Identifier is a standard for an open identifier for organizations introduced by the ISNI International Agency[1]. ISNI identifies public identities of individuals and organizations, is an open identifier and thus may be freely shared with any other person or party without restriction. In this section we propose an algorithm which looks for an ISNI number in an institution string and, if successful, uses it for identification.

[1] http://www.isni.org/content/isni-community

4.3.1 ISNI algorithm

The algorithm for institution identification using the ISNI number is presented in the figure 4.3. The method is triggered only in case when the institution string contains the ISNI number or it is an ISNI number. The ISNI number is a string of 16 characters of which the first 15 are digits and the last one is a check character that can either be a digit or an "X". It can appear as either one sequence of 16 characters or four sequences of 4 characters each.

Examples of the ISNI identification are:

ISNI	Identifing Institution
0000000115037226	University of Porto
0000000403820717	Faculty of Sciences of the University of Porto
0000 0000 9693 350X	University of Algarve
0000 0001 0133 6938	University of Minho

TABLE 4.6: Examples of ISNI identification number.

The Authenticus database stores the ISNI number in the `institutions` table. At this moment there are only few institutions that have this number, but with the usage of this identification method, the column should be quickly propagated.

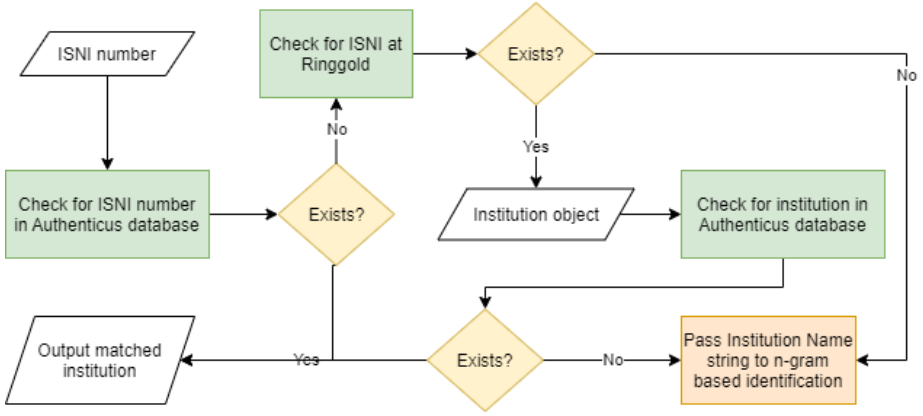


FIGURE 4.3: Diagram of the ISNI based identification method

ISNI Identification

The algorithm starts with checking if the ISNI number exists in the Authenticus database. For this it uses the pandas transformation:

```
institutions[institutions.isni_id == isni_number]
```

where "institutions" is the dataframe to which the `institutions` table was loaded and "isni_number" is the ISNI number to search. If successful, the query returns one institution and the algorithm ends.

In case the ISNI number doesn't exist in the `institutions` table, the algorithm proceeds to the second part, which uses the third party external service for identification. Ringgold's[1] API service, as already mentioned before, is a free service to provide open access to the ISNI Identifiers and data for organizations. The API is a simple 'REST' service which can be consulted using basic a HTTP request containing ISNI number. The HTTP request is
"GET: http://isni.ringgold.com/api/stable/institution/isni_number", where
"isni_number" is number to use for the request.

The API, if successful, responds with JSON formatted data about the institution in the form of an institution object. This institution object returned by the service contains: the ISNI number, the name of the institution, a list of alternative names, locality, postal code, country code and a list of URLs belonging to the institution.

In the next step, the Ringgold object has to be matched to an institution in the Authenticus database. This is done using two parts of the object provided by the API: name, and URL list. First, we match the first URL from the list provided to the end part of any URL in the database with the regular expression `^.*URL\/?$`, where URL is the URL provided by the API. Finally, from those institutions, we apply Levenshtein distance between the name provided and the name of the potential institutions and choose the one with the lowest score (most similar to the name provided).

If a match is successful, we update the local data in the `institutions` table with the new information extracted with the API and return the identified institution in a sequence structure, similarly to what is done in the email identification method.

The identification is unsuccessful when the Ringgold service does not return any institution data, or we are not able to associate the institution object provided by the API to an existing institution in the Authenticus database. In this case, the identification will be forwarded to the n-gram based method.

If the ISNI number is only a part of the initial institution string, the whole string will be sent to the n-gram based identification method. When the institution string is an ISNI number and the Ringgold API returned an object but it was not possible to match it with

[1] https://www.ringgold.com/isni/

any institution in the Authenticus database using the basic matching algorithms, then the name retrieved from the institution object is sent to the n-gram based identification method.

After identification by the n-gram based method, information about the institution in table `institutions` is complemented with the information provided by the institution object.

4.4 N-gram Based Identification

The n-gram based identification method is used when the other methods are unsuited for an affiliation string or the other methods fail in identifying an institution.

An n-gram is a contiguous sequence of n-items from text or speech, in our case, n-words from affiliation strings. Normally n-grams are used in probabilistic language models for predicting the next item in the sequence, but we use them in conjunction with tf-idf values in order to determine word patterns that are distinctive between institutions in affiliation strings. Tf-idf, or term frequency-inverse document frequency, is a statistical metric that takes into account how common the term is for one document versus in how many documents the term is present. This metric is useful to wour work in order to determine which grams are more or less important for distinguishing institutions.

In application to our problem of identifying real-world institutions from the affiliating string, we are using the n-grams algorithm to generate the grams and then, by using their frequencies, we predict which institutions they identify. The n-grams are built on a manually verified affiliation strings dataset described in section 5.1.1.

An example of this would be, for the string ”univ porto, fac med, oporto, portugal”, using 4 as the value for n, the generated grams are ”(’univ’, ’porto’, ’fac’, ’med’)”, ”(’porto’, ’fac’, ’med’, ’oporto’)” and ”(’fac’, ’med’, ’oporto’, ’portugal’)”.

Using the grams we try to understand their frequency patterns in relation to the institutions they identify using tf-idf scores. When generating grams for new strings and comparing the frequency of these between different institutions, we are able to generate a prediction of the institution present in the string.

The method is divided in two main parts: training and prediction. The first part is where we use a dataset with manually verified associations between affiliation strings and institutions, which will be described later in section 5.1.1, in order to generate grams from the affiliation strings and link each one of them to the corresponding institution in an

institution-gram pair. Finally, we create a new dataset with tf-idf values per institution-gram pair. In the second part, we take new affiliation strings and choose an institution based on the previously mentioned tf-idf dataset.

We will first discuss the similar initial stage where the algorithm pre-processes the affiliation string and how the grams are calculated from this string. Then we will describe in detail both parts of the method. A visual representation of this method is presented in figure 4.4, showing both diagrams for training (a) and for predicting (b).

4.4.1 String Pre-processing and Grams calculation

The initial stage for both parts of the method is string pre-processing and grams calculation. In this stage, we start by removing special characters and transforming the string into a list of words. This is done using using the split method from the **re** library with the regular expression `[^\w]`, leaving us with a word list composed only of words with alphanumeric characters. To this list we add the city name and country name present in the `publications_address` table for that string, since these words also have meaningful value in helping identifying an institution.

From this list we remove stop-words, which are normally the most common words in a given language and used in sentences to link more meaningful words together, but don't have much meaning by themselves. For stop-words data, we used the stop-words lists for English and Portuguese available in the corpus library of the nltk package.

Finally, to form the grams, we pass this word list to the ngrams method also provided by the nltk package, which returns a list of grams that will be passed on to the next stages of the method.

These different stages can be shown using the string "instituto de ciencias sociais;universidade de lisboa;anibal de bettencourt". First, we would get the list "['instituto', 'de', 'ciencias', 'sociais', 'universidade', 'de', 'lisboa', 'anibal', 'de', 'bettencourt', 'lisboa', 'portugal']", after dividing the string into words and adding the city and country. Then, after removing the stopwords, the list becomes "['instituto', 'ciencias', 'sociais', 'universidade', 'lisboa', 'anibal', 'bettencourt', 'lisboa', 'portugal']". Finally, using n-grams of size 4, for this example, we would get the list "[('instituto', 'ciencias', 'sociais', 'universidade'), ('ciencias', 'sociais', 'universidade', 'lisboa'), ('sociais', 'universidade', 'lisboa', 'anibal'), ('universidade', 'lisboa', 'anibal, 'bettencourt'), ('lisboa', 'anibal, 'bettencourt', 'lisboa'), ('anibal,

'bettencourt', 'lisboa', 'portugal']". With a bigger n it is less likely that two different institutions have the same n words in their corresponding examples. However, with a bigger n more computational power is required for the functioning of the algorithm. We chose 4 for the n value, since 5 or more would give marginal improvements (less than 0.01 difference in f1-score) while needing significantly more computational power.

4.4.2 Training

Training is the part of the method were we try to learn from previously manually identified strings (section 5.1.1) the frequency of the grams generated for each institution, which will allow us to predict institution in new strings. It is only ran once in order to learn from the dataset. However, if new data is added to the dataset, it can be re-run in order to keep the information updated.

This step starts by iterating through each string of the dataset described in section 5.1.1. For each string we generate the grams as mentioned in sub-section 4.4.1 and update the frequency at which each of them occur related to the identified institution. After the cycle, we obtain the frequency of each gram per institution for the whole dataset.

The next step is to generate tf-idf scores with the frequencies. Tf-idf, or term frequency-inverse document frequency is a statistical metric intended to reflect the importance of a term to a specific document in a collection of documents. In our use case, the documents are the institutions and the terms are the grams.

Finally, this data is stored in a new dataset where each gram-institution pair has a tf-idf score, which will be use in the prediction stage.

4.4.3 Prediction

Prediction is the second part of the method and the one that will be most used in the normal operation of the algorithm for this method. It is where we try to identify an institution from a new string using the data gathered during training to analyze the string.

Similarly to training the first step is to generate the grams from the string as described earlier in sub-section 4.4.1. After this, for each gram, we get all institutions which have some frequency of that gram and normalize their tf-idf scores between 0 and 100. Finally we sum all the scores per institution and choose the one with the highest score as our prediction. The reason we normalize the scores per gram is to insure that, for one institution, a very high score in one specific gram doesn't hide low scores in others.

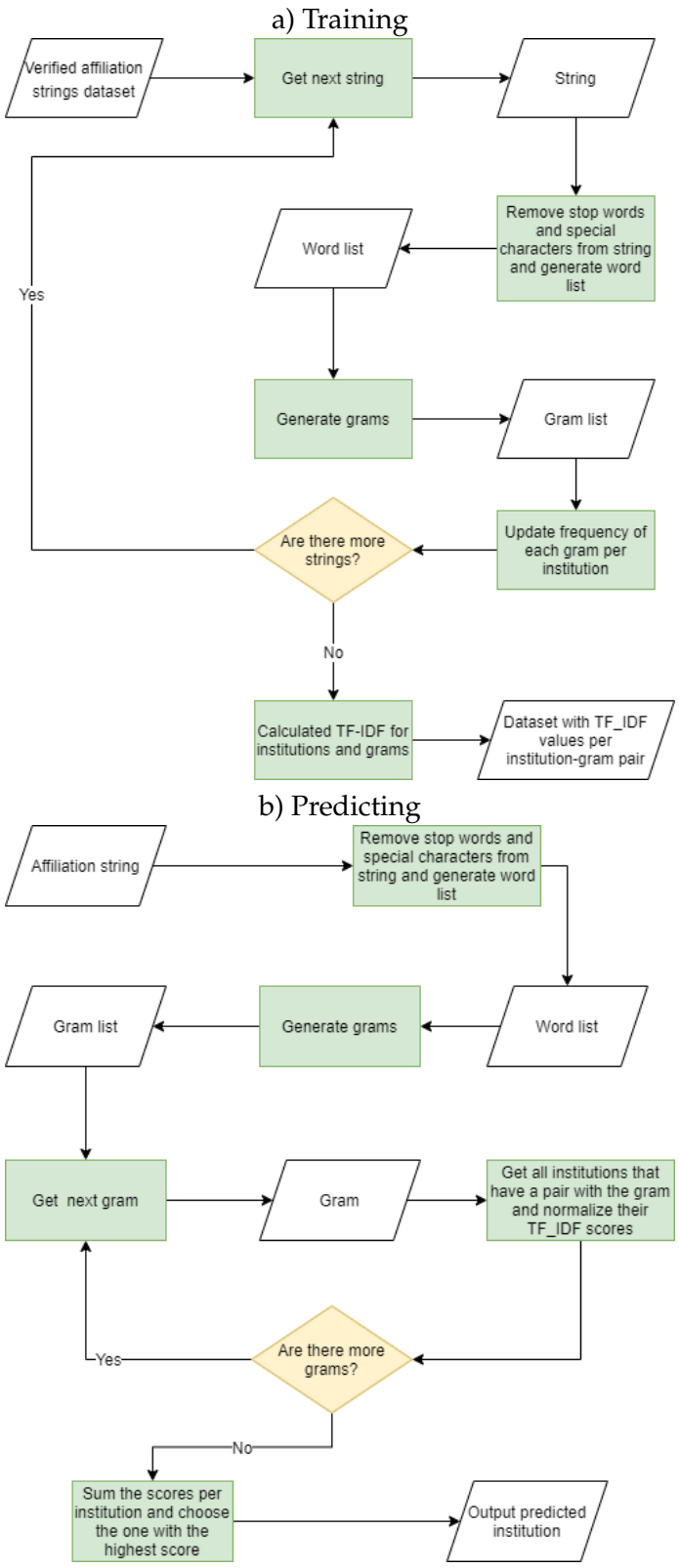


FIGURE 4.4: Diagrams of the n-grams based identification method

Chapter 5

Methods and Results

In this chapter we discuss how we tested our algorithm and present the final results gathered from this testing. We start by describing the datasets used for evaluation and explaining the evaluation method itself. Later we describe several tests performed along our work and we present our results. Finally, we discuss the limitations of the algorithm and describe the API created for the interaction with the algorithm in the experimental case.

5.1 Datasets for algorithm evaluation

For the evaluation of the algorithm, two dataset will be used. The first dataset is one which already existed in the database and was used to evaluate the email identification and n-gram identification methods. The second dataset was generated during our work in order to be able to evaluate the ISNI identification method. These two will be refered as main dataset and ISNI dataset, respectively.

5.1.1 Main dataset

The dataset used to evaluate the email identification and n-gram identification methods is a collection of affiliation strings that have been associated with corresponding Portuguese higher education institutions and/or research centers. The dataset was created in the context of a study "A evolução da ciência em Portugal (1987-2016)"[14], which provides a geography and radiography of the science done in Portugal in the last 3 decades. In that study the Authenticus database was used to process publications data in order to elaborate statistics at regional and institutional level.

This dataset was build from affiliations strings and author's email addresses extracted from metadata of scientific publications published in Portugal in the last 30 years. The metadata was imported into Authenticus from external bibliographic databases, namely WOS and Scopus. It includes over 810k records identified with Portuguese higher education institutions and/or Portuguese R&D centers. The association between affiliation string and institution was made in a semi-manual manner, using SQL queries. The queries were build based on observed by the team rules on how institution are described in the affiliation strings. For example, every time the string matched a SQL condition such as: `affiliation LIKE '% uaberta,%'` the affiliation string was associated with institution `Universidade Aberta`. The results of the queries were verified by the team working on the project.

Data Format The dataset is kept in a table called `final_verified`, and its structure is shown in table 5.1.

Name	Description
pid	Id of the publication
affiliation	Affiliation string
type	Source of the affiliaton string ('wos', 'scopus', 'wos_reprint', 'email' or 'email_reprint')
he_id	Id of identified institution of higher education
rd_id	Id of identified research center
email_id	Id of institution identified via email address

TABLE 5.1: Data structure for table `final_verified`

Data characterization The dataset is comprised of 816594 records, from which 681348 are usable for the evaluation, meaning they have an affiliation string and at least one institution identified. These records identify 333 different higher education institutions and research centers. They are divided into 6 types of affiliations, depending on their source:

- wos - 402758 records (49.3%)
- scopus - 1656 records (0.2%)
- ccips - 36159 records (4.4%)

- wos_reprint - 157589 records (19.3%)
- email - 148660 records (18.2%)
- email_reprint - 69770 records (8,5%)

One limitation of this dataset is that, for `he_id` and `rd_id`, the identified institutions are only of top hierarchical level, meaning that one string containing "Faculty of Sciences, of University of Porto" will only identify "University of Porto", for example. This situation doesn't apply to the cases where the string is an email address and it has associated `email_id`.

Table 5.2 shows the variety of institutions identified by each of the three id types present in the dataset. It includes a count of how many institutions are identified and several metrics about the frequency of these institutions.

For example, for `he_id`, we have 37 different institutions identified. From these, the least frequent one appears 56 times. An average institution appears around 17140 times, with a standard deviation of around 28101 for the distribution. The institution that appears the most times appears 123347 times.

Institution type	**Institution count**	**Minimum Occurences**	**Mean**	**Standard deviation**	**Maximum Occurences**
he_id	37	56	17140.595	28101.104	123347
rd_id	78	1	612.705	1049.102	5199
email_id	281	1	777.338	2331.446	21628

TABLE 5.2: Variety of institutions in the dataset

For a better visual comprehension, figure 5.1 shows violin plots for the three types of institutions. Violin plots are similar to box plots, with the exception that they also represent the probability density at the different values for the data. In our case, the wider the plot is for a certain value of frequency, more institutions are at that value of frequency.

In the figure there are four graphs. The first one shows all three types in the same scale, to facilitate a comparison between them. The last three show individual violin plots for **he_id**, **rd_id** and **email_id**, respectively.

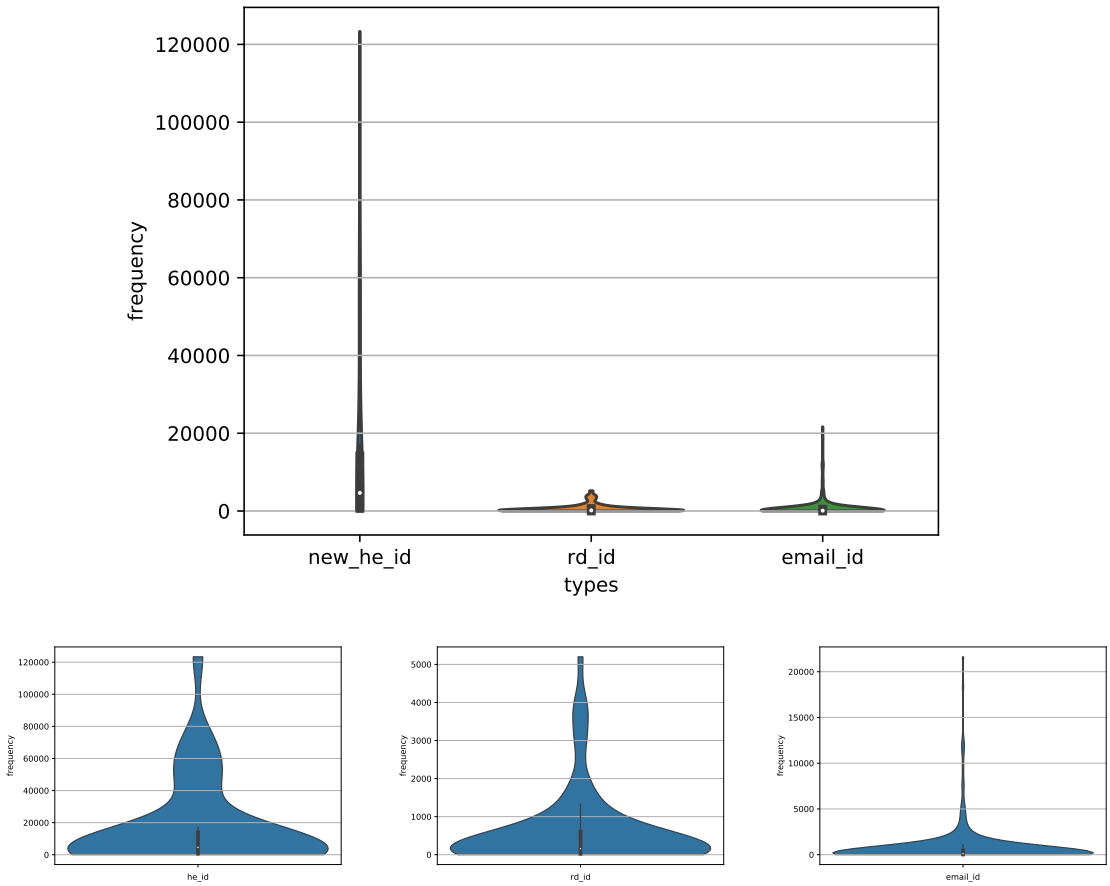


FIGURE 5.1: Violin plots of institutions in the dataset

From these graphs we can conclude that the distribution of all types has more institutions with lower frequency, with **he_id** having more examples with higher frequency compared to other types, and a minimum frequency of 56, where other types have institutions with only one example. In the case of **email_id**, having few examples per institution is not significant, given the fact that only the email domain corresponding to the institution will be used.

5.1.2 ISNI dataset

To evaluate the performance of the ISNI identification method a new dataset was created. To generate this dataset, we selected 100 random Portuguese institutions from the `institutions` table and searched their name in Ringgold's API. The HTTP request used for this search is

"`GET: http://isni.ringgold.com/api/stable/search?q=institution\_name`" where

"institution_name" is the name of the institution to search. The API returns a list of possible results from which we select the first one and retrieve the ISNI number corresponding to the institution. After doing this for all institutions, the dataset is generated as a pair of institution ID and ISNI number.

Finally, we manually verified each entry to make sure the match was correct and, if not, correct the wrong ones.

5.2 Description of Evaluation Method

The evaluation process is crucial to understand the performance of our algorithm. We decided to assess each component of the algorithm separately, with the subset of the dataset corresponding to what these components would try to identify in the normal operation of the algorithm. The evaluation methods and the results of the components, email based identification, ISNI based identification and n-gram based identification are described below.

The main dataset was divided into two parts: (1) strings that are considered email addresses by matching the regular expression described in section 4.2 and (2) remaining strings that contain regular text affiliations. The first was used to evaluate the email identification method and the second for the n-grams identification method. The ISNI dataset was used to evaluate the ISNI identification method.

Email Based Identification

In order to evaluate the email based identification, we decided to classify the method identification results with one of three possible cases:

1. Institution is directly identified, meaning that the institution is the one to which the email domain corresponds to. An example of this would be the email domain "fc.up.pt" corresponding to Faculty of Sciences of University of Porto.

2. Institution is indirectly identified, meaning the identified institution contains the correct institution to which the email domain belongs to, in terms of hierarchy. An example of this case is when for the email domain "fc.up.pt" the algorithm only identifies the higher level institution, namely University of Porto based on the substring "up.pt".

3. Institution is not identified, either because various institutions use the same email domain or because the email domain was not present in the Authenticus database.

When case 1 fails, either case 2 or case 3 will happen. To better understand why direct identification was not possible, the latter cases contain two sub-classifications for evaluation based on the reason why direct identification failed: email domain corresponds to multiple institutions and email domain does not correspond to any institution. The final results are presented in the form of percentages for each case in the given dataset.

ISNI Based Identification

For evaluation of the ISNI based identification method, we decided to classify the identification results as one of three possible cases:

1. Institution was correctly identified.
2. Institution was wrongly identified (false positives).
3. No institution was identified.

The second case occurs when an institution is identified, but the identification is not correct. Analysis of this cases shown that the false identification happens when the data in the Authenticus database of institutions is outdated or wrong. A common case was when one or more institutions have the same URLs, but the correct institution have, either a different URL from the one provided by Ringgold, or no URL in the Authenticus database. The final case is when we were not able to identify any institution.

N-gram Based Identification

The n-gram based identification method outputs a pair of institutions, one of higher education and one research center, where either one can be a null value. Due to the difference of distribution between institution of higher education (`he_id`) and research centers (`rd_id`) the results will be presented separated for each type of institution.

For evaluation we used two scenarios: (1) shuffling the dataset and splitting it with 80% used for training and 20% for predicting and (2) using cross-validation with 5 folds. The second method is similar to the first in the sense that it splits the dataset, however, it splits it in 5 equally sized parts with similar amount of examples of each institution. These parts are then rotated in a way which uses one of the parts for testing and the remaining for training, until every part has been used for testing. In the end the results of each run are averaged to give the final score. The reason for using 5 parts in the second scenario, was due to the fact that, of all the institutions from **rd_id** (78 institutions), 11 had less than 10 examples and 5 had less than 5 examples, and, therefore, a higher division of the

dataset could compromise the results for research centers. For each scenario, precision, recall and f1-score are calculated and used in order to evaluate the performance.

5.3 Tests

In this section we discuss the evolution of the identification methods from the beginning ideas to their final state, explaining the reasons for the alterations along the way. First we will discuss the evolution of the email identification method and later the same will be done for the n-gram identification method.

5.3.1 Email identification

The starting idea for this method was to identify any email address contained in an affiliation string, extract an email domain, identify the corresponding institution and output it. However, this method gave some poor results, which can be seen in table 5.3.

Type of Result	Multiple Matches	Amount	Total
Correctly identified	No	28%	61%
	Yes	33%	
Incorrectly identified	No	0%	7%
	Yes	7%	
Not identified	-	32%	32%

TABLE 5.3: Results of the initial email identification method

Besides having almost a third of cases left unidentified, from the cases with multiple matches for one email domain, the majority (33% of the total) were correct, since the correct institution happened to be the first one on the result order. This is a trend that can be observed in the `institutions table`, where the top institution is usually earlier in the table. However, it is not a rule and should not be considered accurate, which can be also shown by the fact that 7% of the cases were wrongly identified, all of which had multiple matches. Because of this, some changes needed to be made.

The first modification we made was to make the algorithm to be aware of email domains that are used by more than one institution in the dataset. This is due to some institutions using the same domain for itself and all of its sub-units. An example of this

can be shown using the email domain "ualg.pt", which is the email domain used by University of Algarve. However, all of the faculties belonging to this university also use the same email domain, making it impossible to choose one institution to identify without knowledge of the hierarchical structure. The method we used was to compare the email domain with the URLs present for each institution and check if they matched directly, also described in section 4.2. The results from this change can be seen in table 5.4.

Type of Result	Fail Reason	Amount	Total
Correctly identified	-	48%	48%
Incorrectly identified	-	0%	0%
Not Identified	Multiple Institutions	20%	52%
	No Institutions	32%	

TABLE 5.4: Results of first change to the email identification method

This modification removed cases that were correctly identified by chance (of several institutions for one email domain, the correct one appeared first in the dataset) and cases that where wrongly identified. However, it left us with more than half of cases (52%) left with no identification.

Finally, we used the hierarchical feature of the email domain to try to identify the parent institution if the direct identification fails. Hierarchical feature means that one email domain, in some cases, not only identifies one institution but also all its top level parent institutions. As an example, for email domain "dei.fe.up.pt" we can identify a department, faculty and a university. In result, the email identification method was extended with the recursive procedure executed as long as the institution is not identified and the email domain contains enough elements that qualifies for further verification as described in section 4.2.

5.3.2 ISNI identification

Initially this method started as a theoretical method only. We decided to identify an ISNI number using the regular expression already described in section 4.1, when describing the ISNI identification method. With this number we simply would match it to one in the `institutions` table for the correct institution. While no data exists currently on the Authenticus database, with time this data would be added and we would have an immediate correct identification.

The first changed was to add a way to match the ISNI number to an institution, if the number was not present in the database. For this we decided to use Ringgold's API service which can give several data points about an institution with only the ISNI number. We then matched this data to an existing institution using the given URL list and, when several institutions matched the first URL, use Levenshtein distance to compare the given name to the name of potential institutions.

The next change was to, when a match occurs between the API service and the database, to use the information provided to complement the existing one in the `institutions` table, when lacking.

Finally, we decided that, for cases where identification was not possible, the initial affiliation string containing the ISNI number would be passes on to the n-gram identification method.

5.3.3 N-grams identification

This method started as an attempt to identify institutions using word frequency related to the previously identified institutions in the dataset mentioned in section 5.1.1. The first implementation consisted of two parts: a training part and a predicting part. The training part consisted of counting the frequency of each word of an affiliation string and adding it to the corresponding institution, ending up with a dataset with the frequency for each word per institution. The predicting part consisted of processing a new affiliation string, for each word, all institutions with a frequency of said word would be added to the pool of potential institutions. The institution with the highest sum of frequency between all words would be chosen. Both step where executed separately for higher educations and for research centers. However, this method yielded very poor results, which are shown in table 5.5.

Type of data	Type of institution	Precision	Recall	F1-Score
Train-predict split	Higher Education	0.13	0.08	0.06
	Research Center	0.01	0.01	0.01
Cross validation	Higher Education	0.13	0.08	0.06
	Research Center	0.01	0.01	0.01

TABLE 5.5: Results for the initial word frequency identification method

As a consequence we implemented several changes to the procedure.

The first change was to simplify the training step by processing the data for higher educations and research centers at once, and outputting the frequency dataset with a pair of both institutions for each word. This change helped to improve the efficiency of the step, by only needing to run once instead of two times for the generation of the dataset.

The next step was to use n-grams instead of single words. This improved the performance of the method immensely, since with n-grams context of the words is taken into account for the process of identification. We decided to use grams of up to 4 words. While rare, some affiliation strings only have two words, after the pre-processing step described in sub-section 4.4.1. Because of these cases, we decided to also calculate grams with 3 and 2 words for all affiliation strings and only use this data when a new affiliation string requires it. The results of this change can be seen in table 5.6.

Type of data	Type of institution	Precision	Recall	F1-Score
Train-predict split	Higher Education	0.95	0.90	0.92
	Research Center	0.71	0.47	0.52
Cross validation	Higher Education	0.95	0.89	0.91
	Research Center	0.70	0.47	0.52

TABLE 5.6: Results of the first change to the n-gram identification method

While the results were better, there were still some amount of wrong results, due to the unbalance of examples between institutions. To mitigate this, we switched from using frequency of grams as metric to use a tf-idf score for these same grams. The results can be seen in table 5.7.

Type of data	Type of institution	Precision	Recall	F1-Score
Train-predict split	Higher Education	0.96	0.93	0.94
	Research Center	0.80	0.57	0.63
Cross validation	Higher Education	0.96	0.89	0.92
	Research Center	0.70	0.47	0.51

TABLE 5.7: Results of the second change to the n-gram identification method

There were some wrong predictions where the score of one gram was considerably higher than the remaining ones, despite the institution with the higher score sometimes not having even a score for the other grams. An example of this situation can be shown with the string "Univ Minho, CTAC Terr Environm & Construct Res Ctr, Sch Engn, Dept

Civil Engn, Azurem Campus, P-4800058 Guimaraes, Portugal". For this string University of Porto managed to get a score of 816 on the gram "engn dept civil engn", while University of Minho only got 218. However, in the other grams of the string, University of Minho got a much higher score. In the final sum, even having a lower score in most grams, University of Porto was chosen due to the much higher score in that specific gram.

Our mitigation for this issue was to normalize the tf-idf score between institutions for the same gram, between 0 and 100 for the sake of comprehension. With this change, for the previous example, while University of Porto got 79% in the gram "engn dept civil engn" and University of Minho got 21%, in the other grams University of Minho got almost 100% and in the sum of scores was chosen over University of Porto by a very substantial margin.

The results of this change can be seen in table 5.8. The improvement was very marginal in relation to the dataset as a whole, only showing in the third decimal place. However it did fix some obvious wrong matches, like the one mentioned previously.

Type of data	Type of institution	Precision	Recall	F1-Score
Train-predict split	Higher Education	0.96	0.93	0.94
	Research Center	0.80	0.57	0.63
Cross validation	Higher Education	0.97	0.93	0.95
	Research Center	0.80	0.58	0.62

TABLE 5.8: Results of the third change to the n-gram identification method

Finally, a special rule had to be inserted for strings affiliated to Technical University of Lisbon and University of Lisbon. This is due to the former stopping it's activity between 2012 and 2013 and integrating the latter from then on. However, affiliation strings still refer to the former after that date, despite being wrong. In this cases, we change our prediction from the former university to the latter, if the publication in question is from 2013 or later.

5.4 Presentation of Results

In this section we present the results obtained from the evaluation of the email identifcation and n-gram identification methods, following the methodology described in section 5.2.

Email identification

First, we have the results from the email identification method, which can be seen in table 5.9.

Type of Result	Fail reason	Amount	Total
Directly Identified	-	48%	48%
Indirectly Identified	Multiple Institutions	11%	39%
	No Institutions	28%	
Not Identified	Multiple Institutions	9%	13%
	No Institutions	4%	

TABLE 5.9: Results for the email identification method

This method shows some good results and has the added bonus of certainty that successfully identified institutions are correctly identified. We managed to identify 48% of the cases directly and 39% indirectly, with 11% due to multiple institutions matching the initial email domain and 28% due to no institutions matching the initial email domain. With these two cases combined, we get a total of 87% of cases successfully identified.

From the cases we did not manage to identify, the majority (9%) were due to the lack of knowledge of hierarchy between institutions. If this information were to be present, only 4% of cases would be left unidentified. It would also mean that there would be fewer email domains indirectly identified.

ISNI identification

The results of the evaluation for the ISNI based identification method can be seen in table 5.10.

Case	Amount
Correctly identified	78%
Wrongly identified	3%
Not identified	19%

TABLE 5.10: Results for the ISNI identification method

The results show a good amount of cases being successfully identified. While there is a considerably number of cases left unidentified, these ones will be passed on to the n-gram identification method and be identified that way. Finally, the number of cases with

a wrong match is very low, given the fact that it is rare for some institution to have the URL given by the API, but not the correct institution.

N-gram identification

The results of the evaluation for the n-gram based identification method can be seen in table 5.11.

Type of data	Type of institution	Precision	Recall	F1-Score
Train-predict split	Higher Education	0.97	0.93	0.94
	Research Center	0.78	0.57	0.62
Cross validation	Higher Education	0.97	0.93	0.95
	Research Center	0.80	0.58	0.63

TABLE 5.11: Results for the n-gram identification method

The method has very good metrics when identifying higher education institutions, reaching an F1-score of 0.95 when using cross-validation. This are very promissing results and show that the method makes very good predictions

However, when identifying research centers, the method shows modest results, specially in terms of recall. This is due the presence of more institutions to identify with less examples to train for each of them, with 11 of 78 institutions only having less than 10 examples. Another reason is that there are many more examples with a higher education institution and no research center than vice-versa (90% versus 6% of the dataset, respectively). From the wrong cases for research centers, 10% where cases where there was no institution and an institution was predicted. On the other side, 86% of cases where there was an institution but no institution was predicted.

5.5 Limitations of the Algorithm

The algorithm created in this work is divided into three identification components: email based identification, ISNI based identification and n-grams based identification. Each of these components suffer from some kind of limitation. In this section we will discuss those limitations on a per component basis.

Email based identification

The email based identification method is mainly limited by the fact that multiple institutions can use the same email domain and due to incomplete information about the

institutions and their hierarchical structure in the Authenticus database. This limitation, however, can be reduced with time as more and more institutions will use the Authenticus system and update the lacking information in the database providing accurate and up-to-date data. With update of the information, the usefulness of the method will rise considerably.

ISNI based identification

The ISNI based identification method is limited by the lack of examples of affiliation strings which contain ISNI numbers in the Authenticus database, due to the standards novelty and still low adoption rate. For this reason we were not able to properly test the method and judge it's performance, although it seems it could be useful for the future and, with adding of ISNI number information to the `institutions` table in the database, could give immediate accurate identification of institutions.

N-grams based identification

The n-gram based identification method is mainly limited by the dataset it uses for training. We were able to produce some very good results for higher education institutions, while the results for research centers were somewhat lacking. With more and better examples of the latter type of institutions, a better result could be seen for these. Another peculiarity of the dataset was the fact that only top-level institutions, in terms of hierarchy, were identified. This limited us in the way we could present the results without the ability to identify a specific lower-level institution.

Another limitation are edge-case scenarios. This was mainly noticed with Technical University of Lisbon and University of Lisbon. In cases which involved this two institutions, it was common for the former to be written in the affiliation string and correctly identified, however, technically the latter should be identified, since the former ceased it's activity at the time of publishing of the publication. While these cases where few in relation to the dataset as a whole, they were still substantial enough to be seen in the results and require special made rules.

5.6 Experimental Case

This work was done with the purpose of implementing the developed algorithm in the Authenticus project as an API.

The API works with simple get requests of four different types.

The first is for querying any string and the method of the algorithm to use will be chosen automatically. This request can be done as follows:
"`\API_URL/institution?q=String`", where "API_URL" is the URL to access the API inside Authenticus and "String" is the string to be queried.

The three remaining types are for identification using a specific method and can be done as follows:"`API_URL/institution/method?q=String`", where "method" is either "email_domain", "isni" or "author_affiliation", for using email identification, ISNI identification or n-grams identification, respectively.

The response is a JSON object with two elements: institution id and institution name. Each of these elements has as value a list corresponding to the sequence structure mentioned previously. In the case of the n-grams method, the value for each of these is a list with two elements, one for predicted higher education institutions and the other for predicted research centers. When using the general method, an additional element is added to tell which identification method was used.

The API then returns one of three responses:

- Code 200: Successful operation. This response returns a json object with the institution id and the name of the institution identified.
- Code 400: Invalid string. This response means the string sent is not valid.
- Code 404: Institution not found. This response is limited to the email and isni method, and to the general method when the input string is an email address. It means the algorithm could not identify the institution in question.

A documentation for this API was done using Swagger[1], an API documentation tool. The interface of the API is presented in the figure 5.2. This figure contains two images: the general view of the documentation and the expanded view for the ISNI component of the API.

[1] https://swagger.io/

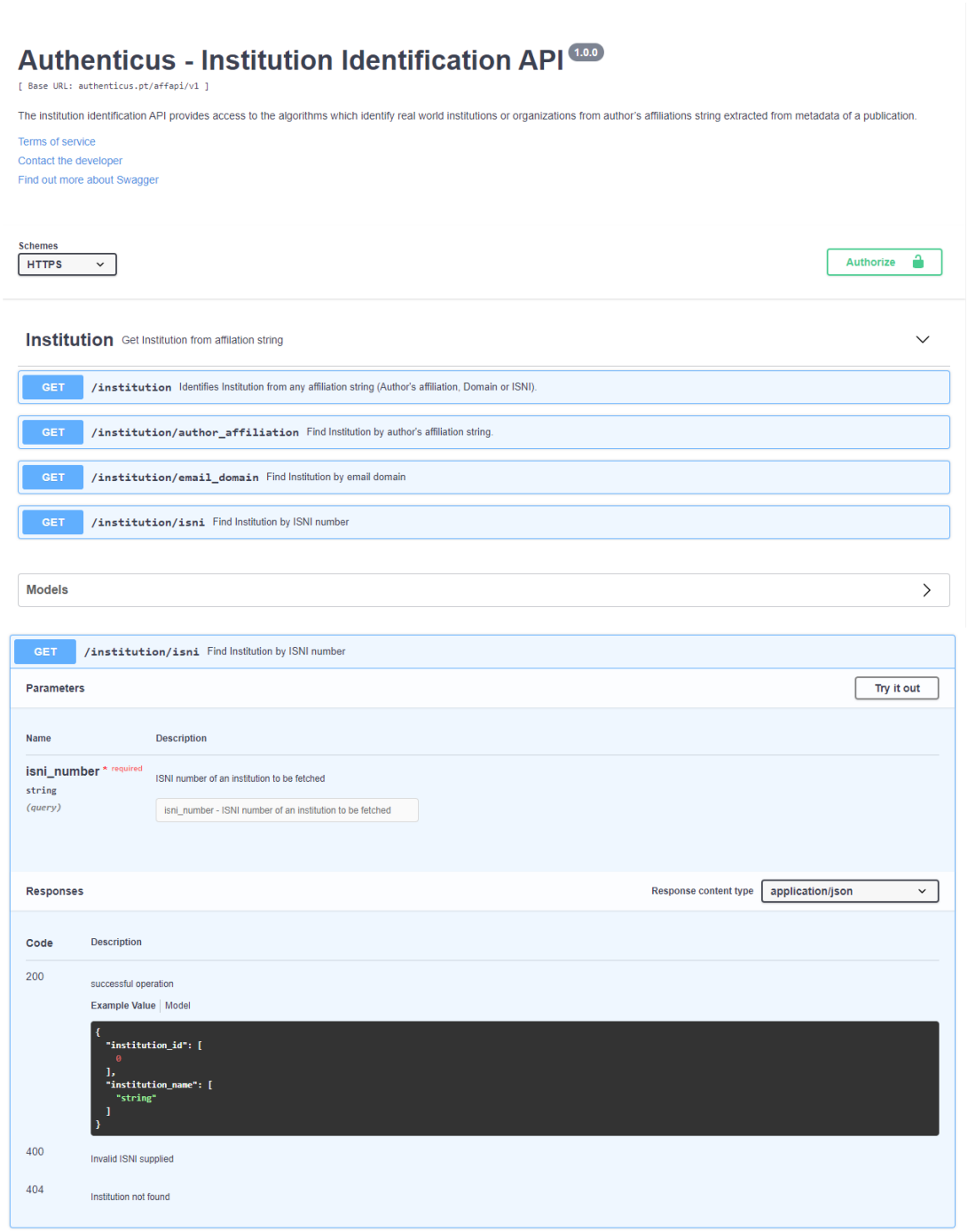


FIGURE 5.2: API interface build with Swagger

Chapter 6

Conclusions

This chapter contains the final conclusions from our work. We start by resuming the work done and describing the main contributions made by it. Finally, we describe some future work that could be done in order to improve what was accomplished.

6.1 Work Description and Main contribution

6.1.1 Work Description

In this document we described an algorithm for automated identification of institutions in affiliation strings.

In order to accomplish this work, we used the Authenticus database to obtain information pertinent to the identification process in several ways, described along the document. For evaluation of the email based and n-grams based identification method, we used a dataset of already verified associations between affiliation strings and institutions, contained in the database. For the creation and verification of the ISNI number identification method we used the Ringgold ISNI API in order to generate a dataset of institutions with their respective ISNI and in order to obtain information not available in the Authenticus database.

The algorithm is divided into three main methods for identification: email based identification, ISNI number based identification and n-grams based identification. One of these is chosen based on the characteristics of the string.

Due to the differences between how each of the methods work, each one has a different evaluation method. The email based identification method showed very good results,

successfully identifying 87% of cases, at least partially, with certainty of the resulting institution. The ISNI based identification method showed good results, correctly identifying 78%, with only 3% of cases wrongly identified. These values should improve as more ISNI data is inserted into the Authenticus database. The n-gram based identification method showed very good results when identifying higher education institutions, with an F1-score of 0.95 using cross-validation, while showing modest results when identifying research centers.

6.1.2 Main Contributions

This document contributes to the studies about institution identification from affiliation strings. These contributions can be summarized in the following:

- An identification method based on email addresses contained in strings.
- A method for disambiguation of the hierarchical structure for institutions with the same email domain using institution URL information.
- An identification method based on ISNI numbers contained in strings.
- An identification method using n-gram and tf-idf in order to identify higher education institutions and research centers contained in strings.
- Creation of a general algorithm that processes strings and chooses the most adequate method for institution identification.
- Development of an API as an interface for the developed algorithm.

6.2 Future Work

The objectives of our work were accomplished by creating an algorithm that successfully identifies institutions in author affiliations with new methods used that are described in section 6.1.2.

While the results of this work were good, some future work can be done in order to improve the performance of the algorithm:

- Obtain more data for research centers identification using the n-gram based identification method, in order to improve it's performance.

- Extend the training dataset for the n-gram identification method to include lower level institutions (currently affiliation strings only match top level institutions) and also international institutions.
- Improve and update the contextual information existing in the Authenticus database, regarding institutions.
- Populate the database with institution's ISNI numbers for better algorithm performance.

Bibliography

[1] C. R. Carpenter, D. C. Cone, and C. C. Sarli, "Using publication metrics to highlight academic productivity and research impact," *Academic emergency medicine*, 2014. [Cited on page 1.]

[2] L. M. Chan and M. L. Zeng, "Metadata interoperability and standardization–a study of methodology part i," *D-Lib magazine*, 2006. [Cited on page 8.]

[3] A. Apps and R. MacIntyre, "Why openurl?" *D-Lib Magazine*, 2006. [Cited on page 9.]

[4] S. Bugla, "Name identification in scientific publications," Master's thesis, DCC - FCUP -UP, 2009. [Cited on page 11.]

[5] T. Poibeau and L. Kosseim, "Proper name extraction from non-journalistic texts," in *Computational Linguistics in the Netherlands 2000*, 2001. [Cited on page 12.]

[6] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, 2007. [Cited on page 12.]

[7] D. Aumueller, "Towards web supported identification of top affiliations from scholarly papers," *Datenbanksysteme in Business, Technologie und Web (BTW)–13. Fachtagung des GI-Fachbereichs" Datenbanken und Informationssysteme"(DBIS)*, 2009. [Cited on page 12.]

[8] W. B. Cavnar, J. M. Trenkle *et al.*, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Citeseer, 1994. [Cited on page 15.]

[9] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the conference pacific association for computational linguistics, PACLING*. sn, 2003. [Cited on page 15.]

[10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, 1988. [Cited on page 15.]

[11] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of tf* idf, lsi and multi-words for text classification," *Expert Systems with Applications*, 2011. [Cited on page 16.]

[12] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (tf-idf)," *ComTech: Computer, Mathematics and Engineering Applications*, 2016. [Cited on page 16.]

[13] D. M. Hawkins, S. C. Basak, and D. Mills, "Assessing model fit by cross-validation," *Journal of chemical information and computer sciences*, 2003. [Cited on page 16.]

[14] E. Vieira, J. Mesquita, R. Vasconcelos, J. Torres, S. Bugla, F. Silva, E. Serrao, and N. Ferrand, *A evolução da ciência em Portugal (1987-2016)*, 2019. [Cited on page 45.]