University of Zurich[UZH]

An approach to automatically gather funding information about scientific research projects from published papers

Bachelor's Thesis in Computer Science

by

Dimitri Kohler
Locarno, Switzerland
12-612-826

Department of Informatics
University of Zürich
Prof. Dr. L. Hilty

Supervisor: Dr. Achim Schneider
November 27, 2016

Abstract

With many parties involved in science, scientific results can be influenced by personal interests. Especially with more and more companies investing in universities and scientific research, the interests of funding entities start to grow in importance. Because of this it becomes easier for companies to follow their commercial interests by influencing scientific results. This can lead to biased results, which can harm the trust the public has in science. To prevent that from happening, transparency about the nature of funding is important. In this thesis it is shown where and in what form funding data can be found and methods to extracting funding information from a paper is proposed and discussed. The two developed approaches use regular expressions and named entity recognition respectively to extract funding entities. Already with a small amount of training data the named entity recognition algorithm performed better than the developed regular expression. The extracted and tagged results are saved in an XML file to be used in further computations.

Contents

List of Figures

List of Code Snippets

List of Abbreviations

AACR2	Anglo-American Cataloging
API	Application Programming Interface
BSR	Behavioral and Social Research Program
CRISP	Computer Retrieval of Information on Scientific Research Projects
DC	Dublin Core
MARC21	Machine Readable Cataloging
NER	Named Entity Recognition
NIA	National Institute of Aging
NLP	Natural Language Processing
NSF	National Science Fund
PDF	Portable Document Format file
RegEx	Regular Expression
SRF	Swiss Radio and Television
TEI	Text Encoding Initiative
TSV	Tab-Separated Values
WoS	Web of Science
XML	Extensible Markup Language

1 Introduction

Today's scientific research is expensive. In order to pursue their research projects scientists must find sources of funding. Universities fund basic scientific research with the goal of education in mind. Public grants give out money to improve the quality and quantity of scientific knowledge in a country or a discipline. The goal of private funders may not only be to expand scientific knowledge but also to pursue other interests. If the funding entity has a commercial interest in the result of the funded science, a conflict of interest happens . From an ethical standpoint funding for scientific projects should not be coupled with other obligations than to produce scientifically correct and objective results. This is not always the case and funding can in some cases influence where scientific research is heading. Big companies fund scientific research through foundations specifically founded to finance science projects that follow the interest of the company. These foundations are often presented as separate entities, which compromises the transparency of the funding and its intent. With biased and manipulated research results the public loses its trust in scientific results. In the worst case science could lose its usefulness and important scientific breakthroughs could be lost. Having transparent funding in science is very important but often neglected. To make a first step to more transparency in the funding of science, this thesis discusses, how scientific funding data of projects and papers can be gathered automatically.

In the first part of the thesis an overview of the landscape of funding in science is presented and possible sources of funding data in the landscape are shown and discussed. Previously developed and used approaches to extract funding data are discussed, compared and evaluated. Based on the evaluation a goal is formulated and possible methods to reach this goal are discussed and implemented as prototypes. These prototypes are discussed and the optimal approach is demonstrated on a small data set. The results of the data set are then analysed and the results visualized. In the outlook ways to further improve and handle the extracted data are shown.

1.1 Intent

Funding data is mainly available in written, unstructured text. Because of this, the data cannot easily be extracted an analysed automatically. This obstructs the transparency of scientific funding and weakens the trust in results of scientific projects. To work with the funding data, it must first be found and extracted from databases and papers that contain this information.

There have already been research projects that analysed the funding flows in science. These projects had to extract the data from somewhere and convert it into a readable form. To develop an approach to automatically find funding entities for a specific scientific paper or project, four already existing approaches to extract, gather and use funding data are reviewed and compared. Based on this review, a goal for this approach is specified that will be used to assess the newly developed approaches in the end.

2 Background

2.1 Landscape of funding in science

A scientific project needs funding to be able to do research. The scientific research produces results that can lead to new knowledge and new discoveries. Figure 1 shows where the funding comes from and where the resulting papers are presented.

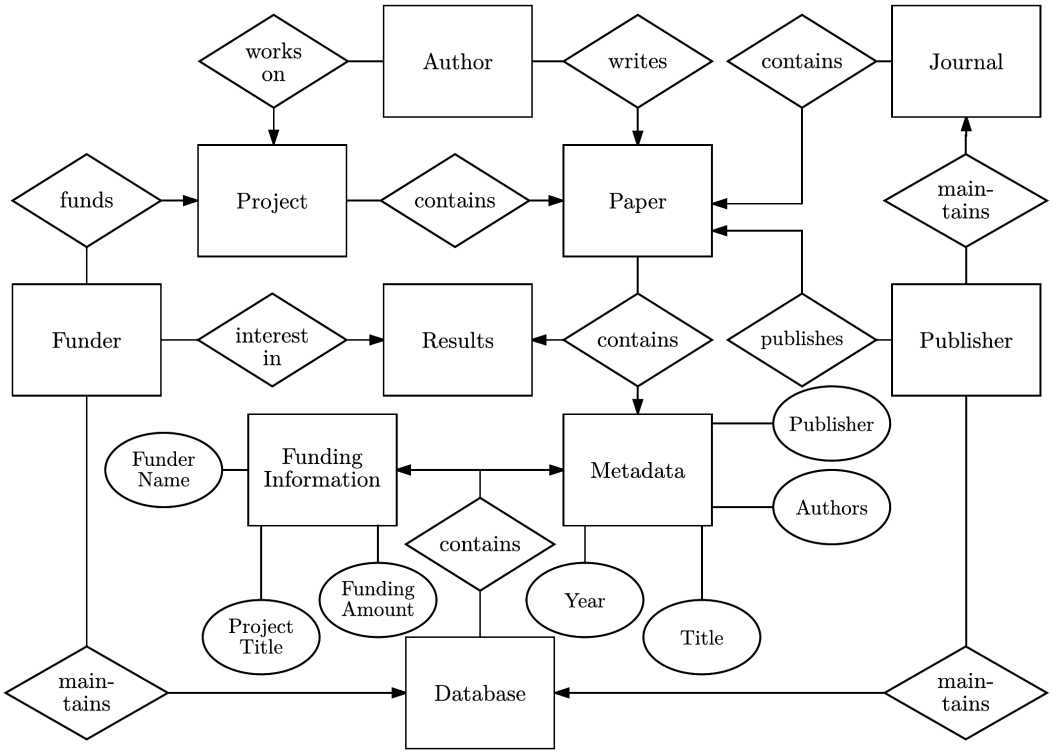

Figure 1: Diagram of the Funding Landscape in Science

A **funder** normally does not fund a single paper directly but invests in a whole project. A scientist with a project can apply for funding from funders such as Universities, public grants or private investors, that are interested in the project's topic and its results. The scientists (**authors**) write papers that bring the project closer to its goal. A **project** is split into multiple papers that cover different parts of the project's topic. The papers are published in **journals** to present the results to the scientific community. Based on these **results** the scientific community can then start new projects to improve the understanding of the topic. The publishers as well as most big funders maintain publicly accessible **databases** that contain information on papers and projects. Databases maintained by funders mainly contain **funding information**, like the funding amount or a project's title. A publisher's database holds paper specific **metadata** but does not contain any funding or project information.

2.2 Metadata of Scientific Papers

Since there is writing there is metadata about these writings. Libraries use metadata to sort books, journals and other writings. With metadata, books could not only be sorted alphabetically but also by their subjects or their year of publication. Where a book was stored has to be documented in catalogues to be able to easily navigate a library.

Today there are different standards for library cataloguing such as the Anglo-American Cataloging (AACR2) [1] or the Machine Readable Cataloging (MARC21) [2]. Modern digital standards for metadata are based on these established standards and rules for cataloguing. Metadata of websites is used by search engines to find relevant webpages for the user's search. To order and search digital writings a standard metadata element set called Dublin Core (DC) [3] has been developed. It defines fields for basic information surrounding a book or paper. DC defines 15 elements that contain relevant metadata. There are elements for the title of the work or the authors name but also for information about contributors, the rights held in and over the source and many more [4]. This Extensible Markup Language (XML) based structure is widely used to describe and store the metadata for books and other writings.

Based on this standard, scientific disciplines have created their own and extended version of DC. In the case of 'cismef', it was used to gather and index health resources written in french to make searching for such information on the internet easier [5] .

What is missing in all these standards however is metadata that gives information about funding. Dublin Core was developed for all kinds of media so it makes sense that it does not contain funding information in the standard. But funding information is also not included in paper/journal specific definitions. Which means funding information is not consistently available in the metadata of scientific papers.

2.3 Publisher Rules

Scientific work is only useful if people can trust the results of the published research and studies. In order to secure the truthfulness and credibility of scientific work, publishers of scientific journals developed rules and ethical standards for papers published in their respective journals. Around 2006 these also started to include guidelines about the declaration of funding and to require an overview of possible conflicts of interest from the authors and the editors of a paper. Even if there is no conflict of interest this fact has to be stated explicitly and in case nothing is declared the paper will be rejected by the publishers [6], [7].

The declaration of funding is handled differently by every publisher. All of them want a declaration to be included in the paper but the rules and format of the declaration are different from publisher to publisher. This fact can make it hard to compare papers from different publishers because the information is represented in different formats and varying degrees of detail [8]. Depending on the journal, the declaration of funding can be heavily regulated in terms of format and detail.

For example, Sage Publishing, which publishes more than 900 Journals in various subject areas [9], writes on their website:

"The funding agency should be written out in full, followed by the grant number in square brackets. [...] Multiple grant numbers should be separated by comma and space." [6]

Sage defines the exact position and naming of the funding section in the paper explicitly. The section needs a separate heading with the title 'Funding' and it must be positioned directly after any Acknowledgement and Declaration of Conflicting Interests and infront of Notes and References [6].

In comparison Springer, another publisher with a wide range of journals, only defines that it has to be declared and that:

"The corresponding author will include a summary statement in the text of the manuscript in a separate section before the reference list, that reflects what is recorded in the potential conflict of interest disclosure form(s)" [7].

The format of the declaration is not defined explicitly but there are disclosure forms provided by the journals to highlight conflicting interests and received funding. Every journal published by Springer can define its own specific standard format, which again can make it harder to compare funding information from different journals.

The amount of money received from a grant and the sponsored entity, i.e. person, project, institution are never explicitly declared in the paper itself. By searching for the grant number in the grants database it is possible to find further information about the funding amount for a project or scientific paper and possibly further information about sponsors of the project. One example for many national and international funds is the US National Science Fund (NSF) that awards grants to science projects and maintains a public database to find information about a funded project [10].

2.3.1 Funding and Acknowledgement Section

The funding section contains text declaring how the project was funded. If authors received individual financial aid from grants or universities, it is stated explicitly. Every funding granted from a public grant has an identification number that is also stated in the funding section. With the grant number, the scientific project can be found on the grant's database.

Today publishers require authors to include an explicit funding section. However, in many cases the funding declaration is still part of the acknowledgments section. The acknowledgments section is similar to the funding section but also includes people and institutions that did not financially contribute to the work.

Because the exact structure of the funding as well as the acknowledgments section is not defined explicitly by any publisher, these sections look different from paper to paper. Thus, they cannot easily be analysed by a computer.

2.4 Funding Information in Databases

Funders as well as publishers maintain databases (Figure 1) that contain further information about the the paper and in some cases more detailed information on the funding of the paper or scientific project. Not shown in Figure 1 are independent databases and search engines that can provide already gathered data about some scientific papers. What all these databases have in common is that they only provide data for a small part of all published papers. Therefore, they are not a consistent source for data on any random paper that could be analysed with the approach developed in this work. They could however be used to cross-reference the results of the initial entity extraction from the paper.

Another problem with these databases is that not all entities in the funding landscape have an interest in making their funding activities public. Most privately owned institutions and foundations do not make their involvement in specific scientific project public. Therefore, they do not maintain a database with more detailed information on the nature of the funding for the supported projects. These funders are often only acknowledged in the paper itself for funding the research with no further information on the details of the granted funding.

In the following sections different public databases are introduced and it is shown how the available data could be used with the results of this approach.

2.4.1 Publishers

Databases from publishers like Thomson Reuters, Sage or Springer contain mainly metadata similar to the Dublin Core standard. Very few publishers gather and publish funding information for each individual paper that is published. This data can be useful depending on the goal of a following work but most probably a publisher's database will not contain funding data.

An example of a database that does contain funding information additionally to the usual metadata for scientific papers is the 'Web of Science'[1] maintained by Thomson Reuters. The search tools give a user the option to search for funding institutions and then lists papers that were funded by the specified institution. The results can be narrowed down further by providing keywords or authors the user wants to look at [11]. However, this works only for papers released after 2008. The service is not free, yet most universities provide access through their network for employees and students.

[1] https://apps.webofknowledge.com

2.4.2 Public Grants

Public Grants maintain searchable databases that mainly contain funding information of scientific projects. Most of these grants are national science and research funds that are financed by a country. These databases give information about funding amount, purpose as well as project title and authors. Without a grant number it is not possible to find funding information for a paper or project.

With a grant number, more detailed information can be found. All public grants that finance science projects maintain their own databases. The databases can be accessed via a grant's website. By searching for the grant number, information about a project and the awarded money can be found [12]. Only public grants make these databases available, most other grants or foundations do not have any publicly accessible specific information about funded projects.

2.4.3 Other Databases

Because the information on papers can be scattered everywhere, various search engines like Google Scholar[2], AGRIS[3] or CiteSeerX[4] have been developed to find scientific work. These search engines also provide automatically extracted metadata. Thus they could be used as universal suppliers for basic metadata. The available metadata has a varying degree of detail from database to database. CiteSeerX in some cases even contains basic funding information but there is no function to search for funding details specifically.

Publishers maintain databases containing scientific papers and their metadata. In Thomson Reuters' Web of Science (WoS) funding information is publicly available in addition to the standard metadata. In the WoS researchers can search for publications and their metadata. The funding data comes directly from the paper or depending on the journal from a funding declaration form the author had to hand in with the paper. In the case of the WoS the funding data consists of the full name of the funding entity as well as a grant number if one is available.

Very few universities have publicly available databases for their publications. Instead, received funding is often disclosed somewhere on a website or in a press release. In Switzerland the Swiss Radio and Television (SRF) Data department has gathered this information from these sources and compiled the data in to a database. The database will not be updated any further but was used to give an overview of the funding in swiss science [13].

[2]https://scholar.google.ch/
[3]http://agris.fao.org/
[4]http://citeseerx.ist.psu.edu/

3 Review of existing Approaches

This work is not the first to use funding data of scientific papers. In this chapter, different approaches are reviewed and evaluated for their strenths and weaknesses. These are then discussed with the goal of finding possible improvements or uses for this project. To narrow it down, four predefined properties will be focused on and rated for every approach. The ranking ranges from 1 to 5 with 1 being the lowest and 5 the highest grade.

1. Internationality:

   How easy is it to use the method in an international context?

2. Generality:

   How general is the approach in terms of different scientific disciplines and publishers?

3. Automation:

   How well is the method automated?

4. Detail of Information:

   How detailed is the information gathered by the method?

The majority of scientific papers are written in English. However, the approach should be able to analyse all scientific work so it needs to be able to process other languages too. If the approach only works for English written papers this hurts the **internationality** of the approach.

There exist many different rules about funding declarations. In order to have a **generally** applicable approach it must be able to find the relevant information in papers of all disciplines and publishers.

The core functionality, extracting funding data from a paper or database, needs to be fully automated. Without a proper degree of **automation**, the approach loses its usefulness and could make the process of finding funding information more complicated.

The **detail of** the extracted **information** is important for further work. With a higher level of detail available, the possibilities and accuracy of further analysis is increased. Examples for higher detailed funding data can be exact information about the amount and the nature of the funding.

3.1 Wang & Shapira (2011)

In their paper from 2011 Wang & Shapira [14] compared Nanotechnology papers on their citation impact in correlation to the source of their funding. They showed that papers funded by national funds generally have a higher citation impact than papers that receive money from other non-government funds. They used data from the global database of nanotechnology publications. This database gathers data on publications in nanotechnology and since 2008 specifically stores funding information for newly added papers. The study looked at national funds from countries all over the word. To compute the data from the database they used a data mining software called VantagePoint.

Because the researched papers are from all over the world it was difficult to fully automate the process [14], but many acronyms and names of different funds often were and still are identical. Many countries for example have a fund called "National Science Fund". Thus, the need for many special cases where it was necessary to manually determine the funding source. For this paper Wang & Shapira did not look at further funding data. No information about the amount or the nature of the funding was considered.

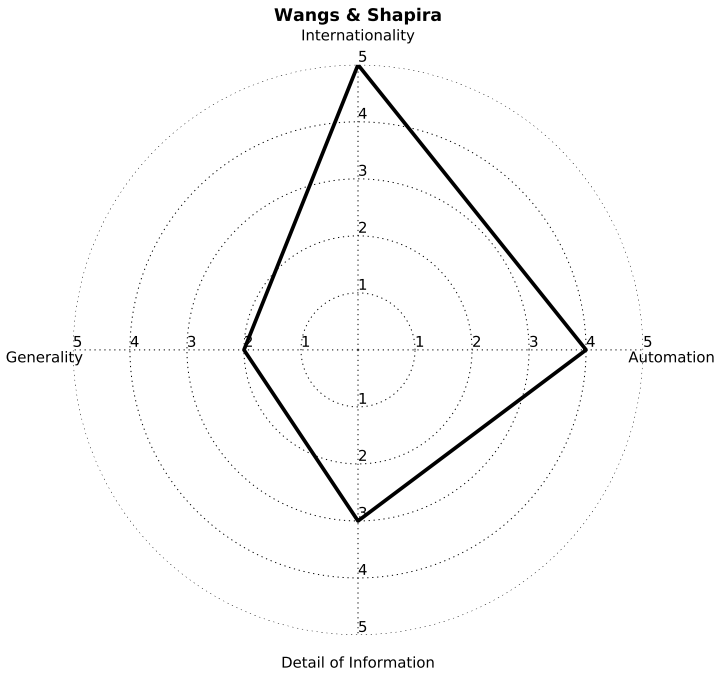

Figure 2: Evaluation of Wang & Shapira's approach

The approach has worked for Nanotechnology because of the readily available database which gathers all the publications but will not work to the same extent in other fields where such a database does not exist. It is nearly fully automated,with a few special cases that must be handled manually. The detail of information is on the lower end. It is however important to note that for the goal of the paper it was not necessary to have more information about the funding.

3.2 SRF Data (2015)

This journalistic approach was used by a team of SRF journalists to gather information about various kinds of funding for projects, institutes and universities in Switzerland [13]. The data was gathered manually from official websites or were directly provided by the universities. However not all universities released their data. The University of Zürich and Luzern did not disclose all their funding upon request and are therefore not part of the database [15]. The data shows amount and purpose of the funding. It was shown that a lot of ethically questionable funding is not always publicly disclosed in all its detail and could potentially influence research and research results.

The data was compiled into a database and visualized for journalistic articles [13]. The Code used to visualize the data as well as the database and the code for the visualization are accessible via a GitHub page[5]. Thanks to this work, Universities in Switzerland are planning on building an official database to make their sources of funding more transparent and more accessible to the public.

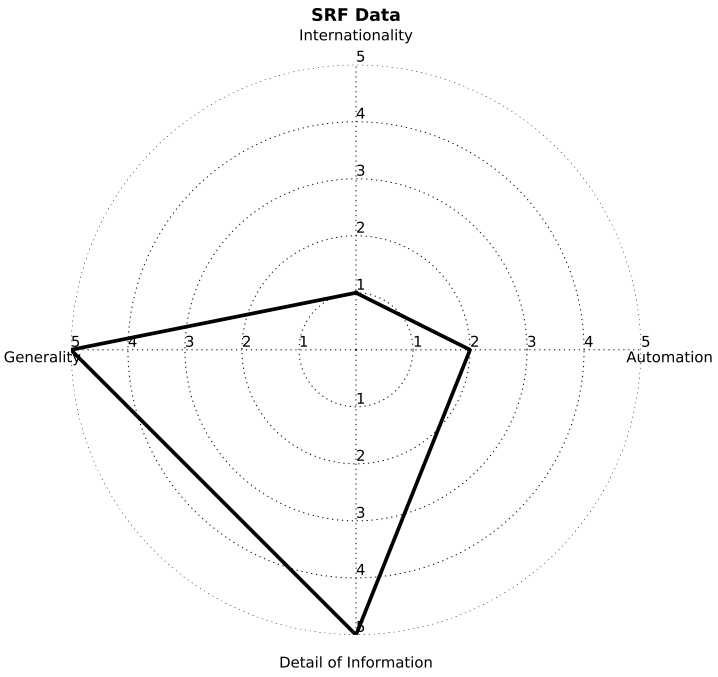

Figure 3: Evaluation of SRF Data's approach

This approach could work internationally, but to use it on an international scale in this form it would require a lot of manual work and be very time consuming. The approach can be used independent from a scientific discipline and is more focused on scientific work at universities. In terms of automation the approach reaches a low score, as finding and extracting the data is not automated. Only the visualization of the data is fully automated. Because the data is handled manually it is possible to make better sense of the data and discern what is relevant. Therefore, the level of detail of the data is very high.

[5]http://srfdata.github.io/

3.3 Giles & Councill (2004)

In 2004, in a time where declaring funding outside of the acknowledgments was not yet common, C. Lee Giles & G. Councill [16] used an approach similar to the already established citation indexing to index the acknowledgments of scientific papers. To index the acknowledged entities, they had to extract the name of the entity from the acknowledgment section in the paper. They found (Regular Expression (RegEx)) to suffice for finding and extracting the data from text. After a lot of tweaking they achieved an accuracy of 98% with their RegEx implementation.

For their work Giles & Councill got papers from CiteSeer (today CiteSeerX[6]) and tweaked the Algorithm/RegEx to index the data found in 335'000 research documents [16]. CiteSeer and CiteSeerX are a project of the Pennsylvania State University which uses citation indexing to search through computer and information science literature.

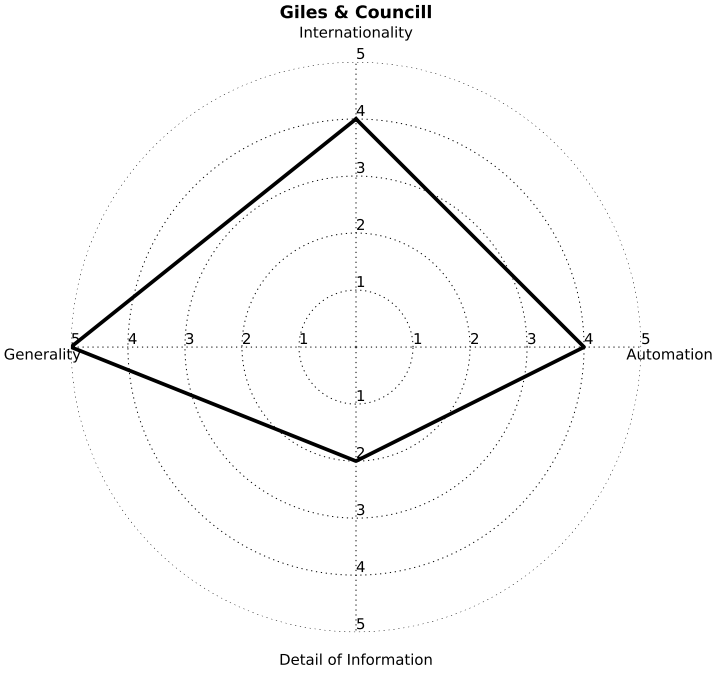

Figure 4: Evaluation of Giles & Councill's approach

Giles & Councill used an approach that can be easily applied internationally. But it is limited to papers that are written in English and indexed by CiteSeer. CiteSeer has indexed papers from a wide range of scientific disciplines so the approach can work in a broad array of disciplines. The data extraction from the acknowledgments section is fully automated, however in 2% of the cases produces incorrect results. The acknowledgments contain very few details, mostly it is just the name of the grant. In some cases, there is information about the grant number but this is wholly dependent on the author of the indexed paper.

[6] citeseerx.ist.psu.edu

3.4 Boyack & Börner (2003)

Boyack & Börner in their paper from 2003 [17] created maps for the landscape of citations linked to grants. The data used to build the maps, was gathered mainly from the National Institute of Aging (NIA) and the Behavioral and Social Research Program (BSR) accomplishment reports. The data from NIA already contained information about grant numbers, awarded amount and standard metadata like title and author for projects between 1975 and 2001. The data provided by BSR had to be parsed so it could be combined with grant data in a single database. More grant data was gathered from Computer Retrieval of Information on Scientific Research Projects (CRISP) an online database which contains data about federally funded biomedical research projects [18]. This newly created dataset was then evaluated and maps were built with various algorithms and programs created by Boyack.

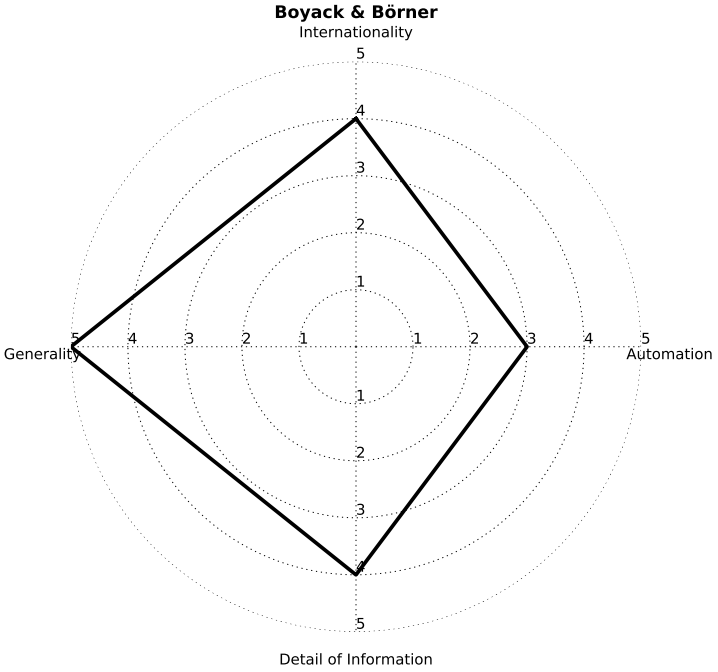

Figure 5: Evaluation of Boyack & Börner's approach

In an international setting this approach could work. But it is completely dependent on the institutions that deliver the data and requires them to have gathered the data in some form. Everything hinges on the data supplier, be it the generality or the level of detail. Because gathering the data is not done as part of the approach itself, all properties, except for the automation, can change depending on the supplied data. In the case of the NIA only few disciplines are covered but the level of detail in the data is high. With other sources for the initial data this may vary. The approach automates the indicator-assisted evaluation and the creation of maps. What lowers the automation score is that it is not shown how the provided data is combined with the grant data nor how the grant data was gathered from CRISP.

3.5 Evaluation

The four approaches have different points in the landscape where they gather the funding data. Wang & Shapira were able to use a database that already had the funding data available in a clean and usable format [14]. This method is not usable in a newly developed and fully automated approach because not every scientific discipline has a database in place to gather all the published papers and store funding information on the papers. The desired degree of generality could not be reached by using their approach.

SRF Data built their own database by manually gathering data from news articles, websites and universities [13]. Because all the data was gathered and reviewed manually this method does not work for an automated approach in this form. What could be interesting however would be to fully automate the data gathering by web scraping news articles and university websites for funding data. This way a high level of detail can be achieved and depending on the implementation a high automation. If the goal is to get project specific funding data, this approach does not work. The data gathered by SRF Data does not contain specific information about single projects but more general data about funding of universities and their departments.

Giles & Council extracted the data directly from papers by using RegEx [16]. This is an interesting approach that could be used and improved upon. Their approach is already well automated and the extracted data could be used to find more detailed information on the internet which would lead to a high level of detail. With newer papers having funding sections that are separate from the acknowledgments the extracted data is potentially less noisy. With less noisy data the funding of a paper can be tracked more easily.

Boyack & Börner got funding data directly from the NIA but had to review and restructure most of it in order to be able to use the data for their work [17]. How the NIA gathered the data that Boyack & Börner received is not known. It seems like it is internal data of funding distribution and projects, that is not accessible publicly. The NIA data is very detailed which makes the level of detail of their approach's results high. This approach requires a source for the data that can provide already structured funding data.

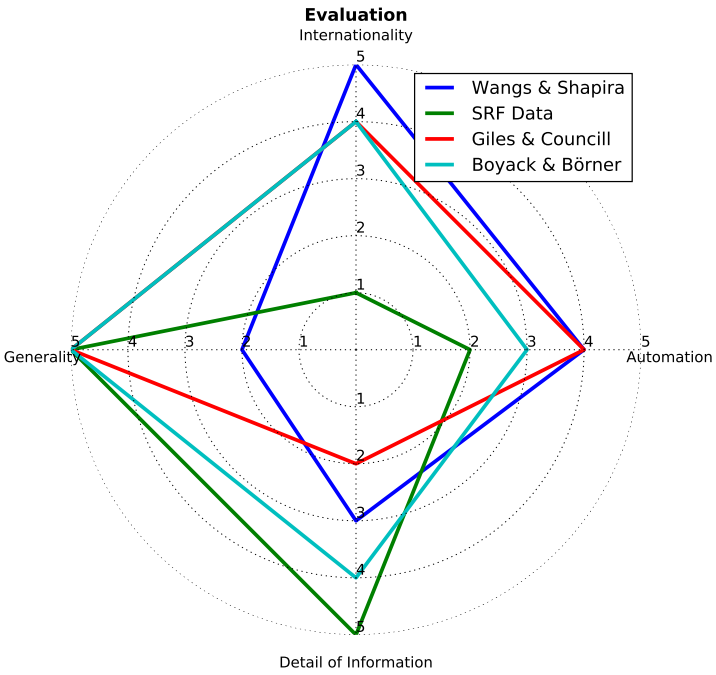

Figure 6: All evaluated approaches in comparison

When Figures 2-5 are laid over each other (Figure 6) we can see that internationality and generality are covered very well by the majority of these approaches.

SRF Data's [15] method gather information on a much higher level of detail than the other three approaches, however the gathering of data is very poorly automated. In Figure 6 it can be seen that no approach manages to have a high grade of automation in combination with a high level of detail. With a higher level of detail the degree to which an approach is automated decreases and vice-versa.

The approaches from Wang & Shapira [14] and Boyack & Börner [17] both cannot be generally applied for all scientific disciplines. Not every discipline has a database or research program that tracks every publications funding. The SRF Data approach is very general however extracts data that is not project specific. With no project specific data, a big part of the SRF approach cannot be applied in an approach to gather scientific paper specific information. Giles & Councils [16] approach is used for to extract acknowledgement data from papers from various scientific disciplines. It has a high automation and could be adapted to extract funding data.

3.6 Goal

Based on the evaluation it became clear, that no approach reaches a high level of detail in combination with a high automation. The goal for this approach is to be able to automatically extract funding data with a higher level of detail than other automated approaches have reached. The approach should fulfill the properties shown in Figure 7 with a focus on the connection between automation and level of detail. This would make finding funding information about a particular paper easier and could enable the public to easily find funding information for science projects.

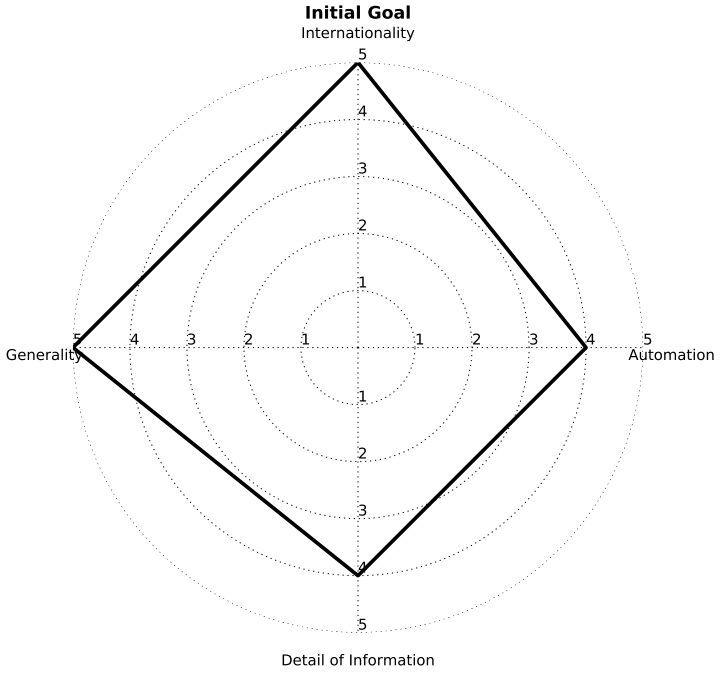

Figure 7: Properties of the goal approach

4 Development of the Method

In order to extract the needed funding- and metadata there are two sources where this data is available. The data can be gathered by either directly accessing a publisher's database or by analysing a papers content and extracting the relevant data. There are also services like CiteseerX[7] that extract the metadata from papers and make this data available and searchable for the public. Because of the vast amount of scientific papers these services are not able to index all published papers. Therefore, they are not a reliable source for the needed funding data. In conclusion the paper is the most reliable source for funding data. To use papers as a source, the approach needs to be able to extract the funding entities from a text and save these results in a machine-readable and structured form that can be used for further work. The following sections describe possible ways to implement this process and provide the basis for the approach demonstrated in chapter 5.

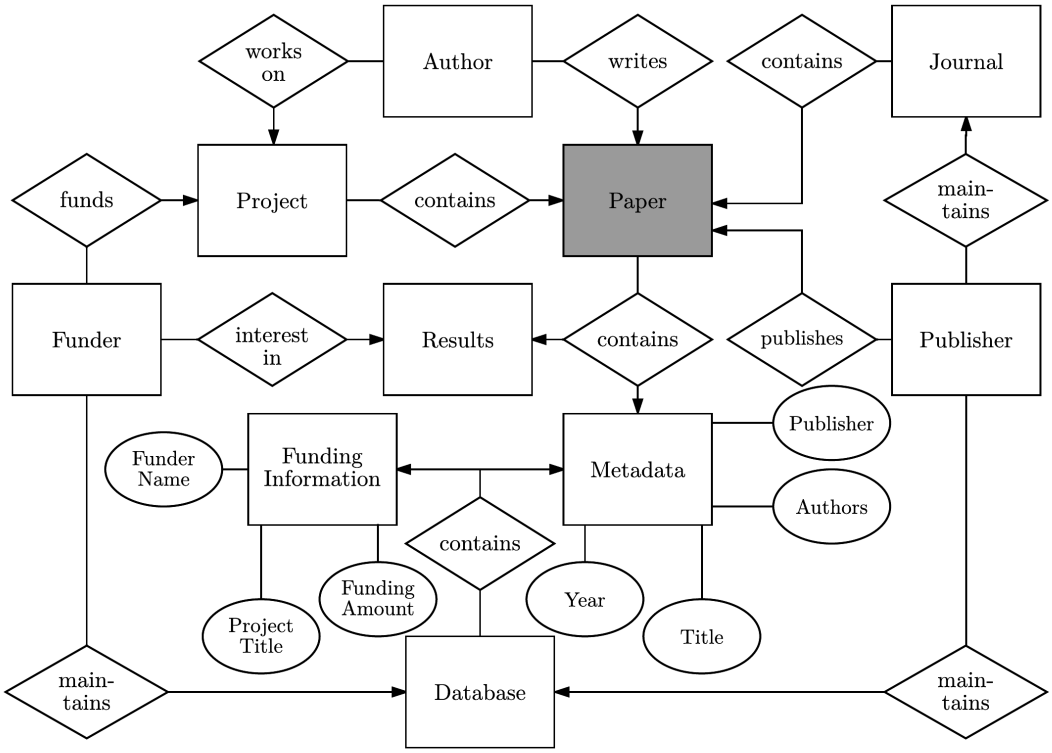

Figure 8: A scientific paper in the funding landscape

Giles & Coucill's approach from 2004 already uses the paper as the central source for the data extraction [16]. Their approach is used to index and extract acknowledgement data. It could be adapted to extract funding data and improved so it could gather the data with a higher level of detail. The new approach could meet the set goal.

[7] citeseerx.ist.psu.edu

4.1 Gathering Data from the Paper

In their paper Giles & Council describe the technologies they used for their data extraction as follows:

"Several approaches have been proposed for automatic metadata extraction, with the most common tools including regular expressions, rule-based parsers, and machine learning algorithms. Regular expressions and rule-based parsers [...] can perform acceptably well if data are well behaved. Machine learning techniques are generally more robust and easily adaptable to new data. [...] Because of recent success using [support vector machines] SVMs for [machine] learning [...], SVMs are becoming increasingly popular tools for classification. [...] While highly effective at metadata extraction, much recent work using machine learning [...] exploits the semistructured format of document headers for chunk identification and classification. The problem of acknowledgment extraction involves the identification of chunks of a single class found most often within free text. We have found that regular expressions work acceptably well [...] within identifiable acknowledgment passages."[16]

They considered the use of machine learning, rule-based parsers and regular expressions, but ultimately settled for RegEx and achieved an accuracy of 98%. Since 2004 machine learning has evolved further and the techniques have been refined. Therefore, Giles & Council's approach must be reconsidered and adjusted for the newly developed approach. In the next sections regular expressions and Machine Learning are discussed, implemented and the results compared to find the best technique to extract funding data from a scientific paper. Modern rule-based parser mostly work with some sort of machine learning, therefore the extracted results would be similar to the results the machine learning implementation achieves.

4.1.1 Regular Expressions

A regular language allows finding simple structures and patterns in strings. A regular language is defined by constructing a pattern using regular expressions. This pattern is then matched with the text and returns all matches. For the English language, a pattern can be created to match entities. The wanted funding entities all start with an uppercase letter followed by lowercase letters and one or more uppercase word separated by a white space. Code Snippet 1 shows this principle used in a regular expression.

Code Snippet 1: A basic regular expression

```
([A–Z][a–z]+\s?){2,}
```

This expression can be used to match different entities. It can match names like "Dimitri Kohler" or names of companies and institutions like "Salus Mundi Foundation". What it does not match are single uppercase words, lowercase words and numbers. It also will not return results containing any other signs than letters. Not all entities have a name in a easy to recognize structure like the "Salus Mundi Foundation", some entities contain lowercase conjunctions like the "University of Zurich". In order to extend the recognized structures, it is

necessary to expand and tweak the expression. The developed regular expression will only work for English text. It will struggle with other languages because they either other letters and signs (i.e. Chinese) or have a different use of capital and small initial letters (i.e. German). For other languages, the expression would have to be rebuilt with the relevant signs or with a new concept.

4.1.2 Machine Learning

Machine learning is different from the previously described method. It can be used for a wide range of tasks, like image recognition or economical predictions . In this case, it is used to learn and recognize (funding) entities [19]. This field of machine learning is called Named Entity Recognition (NER). Entities are recognized by learning the structure of different sentences from training data. With the training data, the machine learning algorithm can train a model.

A model is a collection of rules and probabilities and can come in different forms like a decision tree or a dictionary. To extract entities from text the machine learning algorithm applies the model and returns the results with the highest statistical probability or only exact matches, depending on the implementation. The model learns the structure of sentences and how it can recognize the in the training set defined entities. For the learning process it is not only relevant how the entity itself is structured but also what words and signs are written before and after the entity. It learns from the context the entity appears in and can then recognize a different entity in a similar context [20].

Therefore, the training data must contain full sentences or sections and not just the entities by themselves. In the training data, every word needs to be tagged with its entity [20]. Most machine learning implementations need files in a Tab-Separated Values (TSV) format [21]. The TSV file contains tokenized text which means that every word and punctuation is put on a separate line. Every word or punctuation is then tagged with the according entity. If a word is not part of an entity it is tagged with a 0 or a similarly unique character depending on the implementation of the machine learning. A perfectly tagged training set sets the gold standard for the entity recognition. In the beginning all training data needs to be manually revised to get the best learning results. After the machine learning algorithm has been trained sufficiently it can create its own training sets [22]. This automatically created training set is not in the gold standard but it can still be used to improve the model.

A big advantage over regular expressions is the possibility to categorize the results. This however depends heavily on the quality and size of the training data. If the initial training data is not reviewed properly the resulting model will be faulty which leads to wrong results and wrong categorization. A small training set should not be used to train a model on a big number of different tags. In the case of funding entities, the main focus is on the differentiation between persons and organizations. With a very big training set, organizations could be split further into grants, universities and other sources. This would make the resulting data more detailed and easier to analyse automatically in a next step.

Many implementations of named entity recognition are focused on biochemical work, i.e. to recognize names or abbreviations of genes and substances. These implementations usually come with huge pre-trained models. Such models do

not exist for funding data specifically there are however implementations that use general models to categorize the recognized entities into categories like person, date or location. One such tool is called GROBID NER [21] and is a module of the GROBID open source project. The GROBID application is used to extract metadata from scientific papers. GROBID was coded in java and provides a REST Application Programming Interface (API) to communicate with other applications [23]. It performs a header extraction on a paper's Portable Document Format file (PDF) and can not only extract the traditional metadata defined in the Dublin core standard but also more detailed data like the affiliations of the authors. The metadata extraction achieves an accuracy of 99% according to the developers [23].

GROBID NER is a separate module that can be used to categorize entities in Wikipedia articles or news stories [21]. It comes with two huge models that have to be retrained with their initial training data. The training data has to be requested from Reuters separately and is not available publicly. Because of the size of the training set the training process can take several days depending on the computing power of the computer. The trained tool can categorize the results into 26 hard coded tags [22]. However, most of these tags are not useful in terms of funding data extraction which leads to a large overhead and noisy results. In order to get cleaner results and a more efficient computation it is necessary to be able to define custom categories.

As part of the Stanford Natural Language Processing (NLP) project [20] a named entity recognition algorithm has been implemented. The application is called Stanford NER and contains simple models for different languages [24]. The simple pre-trained models use the tags person, organization and location. With a custom training set these categories can be expanded or changed completely. Compared to GROBID NER the training process is faster and easier to handle which makes Stanford NER the better implementation for situations where the model has to be retrained and customized very often. Stanford NER has been programmed in java but has since been ported to other languages or at least provides an interface for other languages to use its functionality [24]. The results are tagged with XML tags and can be saved to a well-formed XML rather easily.

4.2 Handling the Results

The structure of the data after the extraction with one of the shown methods is very simple. For every extracted entity, the program creates exactly one string. These strings can be handled however needed. If the goal is to get information about a single paper it makes sense to save the entities into an XML file possibly together with other metadata of the paper. If this is done for multiple papers it would make sense to save them in a database. This way it would be possible to start a large-scale analysis of the data. In the next two subsections, the structure of the output XML is described and a structure for a database is proposed.

4.2.1 XML

Because RegEx cannot detect different entities, the XML file created from the results, only contains a root element called "FUNDING" containing child elements called "ENTITY". The NER approach can recognize different entities and tag them accordingly. Depending on the training data different tags can be in the resulting XML file. The standard model Stanford NER uses, is only trained with "ORGANIZATION", "LOCATION" and "PERSON" tags. This results in an output file consisting of the root element "FUNDING" containing child elements according to the trained tags. There can be multiple elements of the same tag.

The resulting XML could be combined with results from other tools to expand the available metadata. For example, the output XML file from a GROBID header extraction could be combined with the funding entities of a paper. The resulting file would contain all extracted funding entities and general metadata like author, publisher or title of the scientific paper. The structure of the GROBID output XML is well documented [22] and follows the Text Encoding Initiative (TEI) standard to make the file easily computable [25]. From the XML file the data can either be used in further computations or the data could be saved into a database.

4.2.2 Database

A database could also be used to store all the gathered data. If a big amount of papers is analysed with the approach it is necessary to store the results in a database because it is easier to maintain than multiple XML files. The database can then be used for data mining or big data analysis. The structure of the database depends on the structure of the results. In the case of funding entities, a possible structure could look like Figure 9. The proposed structure has similarities to the funding landscape. It reuses the central objects “paper”, “project”, “funder” and “author” with similar properties.

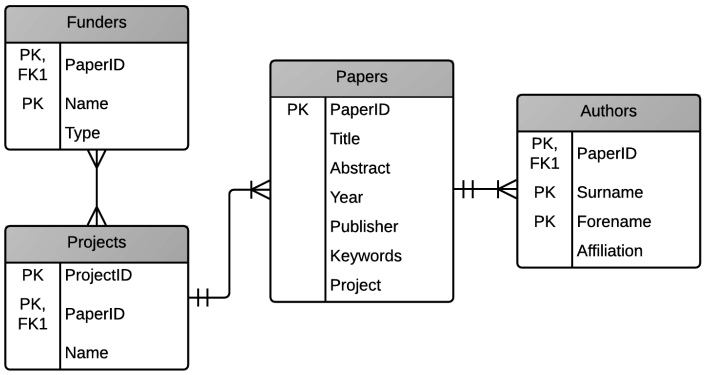

Figure 9: Proposed database structure to save the extracted data

The prototype developed in this thesis will not include a database because a database is not part of the approach and depends entirely on the use case of the approach.

5 Implementation of the Method

In this chapter the developed methods are implemented, tested and discussed. The prototype was coded in java and can be found in the Appendix and on GitHub[8]. In order to test and train the prototype 100 PDFs containing funding information in their acknowledgment or funding sections were used as data. AGRIS[9] and Elsevier[10] were used to find and download the papers. The papers were selected for their simple structured funding texts with clearly recognizable entities. With these 100 papers the NER model was trained to recognize entities.

Before the methods can recognize any entities, the first challenge is to extract the text from a PDF and then find the relevant section, containing funding data, from it. After the relevant parts are extracted from the text, regular expressions and machine learning are used separately to find the funding entities in the text. In the last part of this chapter the results and implementations of RegEx and machine learning are compared and evaluated.

5.1 Text extraction

A RegEx or machine learning algorithm cannot be used directly on a PDF. The content of a PDF has to be extracted before the text can be analysed for entities. Because a paper does not entirely consist of funding declarations it is also necessary to identify relevant passages. In newer papers these passages are clearly labeled with the title "Funding" in older papers the "Acknowledgement" section can be used. There is normally no other section in the paper that contains funding information. This section discusses the tools used to do these two tasks and how they are used in the approach.

5.1.1 Extracting raw text from a PDF

A PDF file is a relatively unstructured way to save data. The data contained in the file is used to draw text and other content on a blank page. This entails that the orders of the text data in the file and the drawn text does not have to be in the same order. Extracting text from a PDF can therefore not be done consistently for all PDFs. There are different implementations of PDF extractors which all handle the data differently. One such implementation is PDFBox[11] which could extract the text from 89 out of the 100 used PDFs while keeping the intended and extracted structure the same. In order to increase the number of correctly read PDFs, PDFBox is supported by PDFxStream[12], another implementation of a PDF extractor. If a PDF's extracted text is found to be faulty PDFxStream is used to try the extraction with another implementation. This way the approach could correctly extract and use 97 out of the 100 PDFs. The 3 PDFs that could not be read properly were built incorrectly and their text could not be extracted in a readable form. These PDFs would have to be rebuilt from the original source text in order to work.

[8] https://github.com/dikohl/funding-extractor
[9] agris.fao.org
[10] https://www.elsevier.com/
[11] https://pdfbox.apache.org/
[12] https://www.snowtide.com/

5.1.2 Finding relevant Sections

After a successful text extraction, the relevant passages containing funding information have to be found in the resulting string. Finding the relevant sections in the text is important because the full text contains entities that are not funding specific. If the entity extraction were performed on the full text the resulting data would be very noisy and the funding entities could not be distinguished from all the others. The relevant funding entities are normally only contained in the "Funding" section and for older papers in the "Acknowledgment" section. Extracted data from the latter can still be noisy because it also contains non-funding entities. These can not be recognized automatically and have to be filtered by hand if the acknowledgements section is analysed.

In order to find the correct section a regular expression was developed that recognizes the word "Funding" and the subsequent text of the section. To find the end of the section the expression uses the fact that the next section will start with its own title (Code Snippet 2).

Code Snippet 2: Simple RegEx to find the funding section

```
(Funding|Acknowledgment)(\n+.*?)+\.)(\n*[A–Za–z]*(\s[A–Za–z]*)?[\n]+
```

Giles & Coucill [16] analysed the acknowledgement sections of papers and used a machine learning algorithm to find the relevant parts of the paper. This method is not used in this approach because it is costly in terms of computing power. Additionally, a regular expression is easier to translate into a different language than creating a new training set for a machine learning algorithm. The regular expression in its current form cannot handle spelling errors and will not find the relevant sections if their titles are faulty. The regular expression has however been tweaked to recognize variations of the basic titles "Funding" and "Acknowledgement" in order to be able to be used on a wider range of papers. A machine learning algorithm would not be dependent on the titles of sections of the paper and could recognize the relevant passages even with missing or misspelled titles. Spelling errors and papers without declared funding sections are relatively rare therefore a machine learning algorithm would not improve the quality of the extracted entities significantly.

5.2 Entity extraction

To extract funding entities two methods using regular expressions and machine learning respectively have been implemented. In the next chapters, the two implemented methods are described, discussed and their accuracy is compared.

5.2.1 Regular Expressions

After the text extraction a string with the funding or acknowledgement text is matched with a regular expression. The RegEx engine that comes with Java does support the basic features of regular expressions to match string. However, it cannot perform look ahead and look behind operations [26]. But there are very few cases where this functionality would be useful. Thus it would not increase the quality of the extracted data.

Based on the expression in chapter 4.1.1 the RegEx was extended and tweaked to be able to match more complex entity names (Code Snippet 3). In addition to uppercase words it now matches entities containing linking words and conjunction words like "of","and" or "-". In order to also get possible grant numbers in the text, the expression extracts brackets that follow immediately after the recognized entity. Because funders can also be private persons their names have to be recognized too. Names are often written with an abbreviated first name followed by a dot and the full last name. This has been implemented by matching a dot after the first uppercase letter.

Code Snippet 3: Regular expression from the prototype (Java)

```
([A-Z](\\.|[a-z\\/]+)[\\s\\-]?
(of\\s|and\\s|for\\s|in\\sthe\\s)?){2,}
(\\s?\\([^\\)]+\\))?
```

A problem with this approach is that it can only handle papers written in English. It struggles as soon as the paper is written in any another language or even if only a single entity has a foreign name containing special characters. To apply RegEx on a paper in another language the whole expression has to be rewritten and adjusted for the new signs and rules. The output of this implementation can be used to build a simple training data set for the machine learning algorithm. The data set will not meet gold standard data but it can be corrected manually.

5.2.2 Machine Learning

For the machine learning algorithm the Stanford NER implementation is used. GROBID NER is not utilized because of the hard coded tags and the fact that it could only be run after several days of training. Because the pre-trained model of Stanford NER performes worse than the developed RegEx implementation on the funding data, a new model is trained specifically for funding entities. The training data was created semi-automatically from 100 PDFs by using the developed RegEx implementation to recognize entities. The recognized entities were marked as “ORGANIZATION” in the TSV file and manually reviewed to achieve a gold standard training set. The created training data and references to the 100 source papers can be found in the Appendix. The new model uses the same tags as the pre-trained Stanford NER model, “PERSON”, “ORGANIZATION” and “LOCATION”. More tags could be added to the model but the training set would have to be expanded. The full gold standard training set used for this work can be found in the prototype of this approach. Because the training data is relatively small, training sets created by the NER implementation should not be used without manual revision yet. To build a training set for another language Stanford NER provides some general pre-trained models for different languages. These could be used to build a language specific model to recognize funding entities similar to how the RegEx implementation was used in this case.

6 Demonstration of the Method

The implemented approaches are applied on two papers and the results are shown and compared. After the funding entities are extracted, the accuracy of both approaches are compared and the approaches are discussed further.

6.1 Entity Extraction

The first paper by Hallin & Briggs [27] has a well structured funding section, with one very long entity name at the end. In comparison to the first paper, the second paper written by Sarchielli et al. [28] has a more complex funding section structure. It contains many different special characters that can be hard to process for both approaches.

6.1.1 Well Structured Funding Section

This paper's funding section is well structured and is the best case for a entity extraction. The contained entities are split very clearly and the text does not contain any special characters. The RegEx approach was built for this kind of funding section and the NER approach was trained with similarly structured funding declarations. The next examples (Code Snippets 4 - 5) shows how outputs from the RegEx and NER implementations look like and what the gold standard for these results would be. The initially analysed funding text looks like this:

"Funding This article emerged from research funded by the Salus Mundi Foundation and supported by the University of California, Berkeley and University of California, San Diego. It grew out of discussions that took place while Daniel C. Hallin was a Fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University."[27]

Code Snippet 4: Gold Standard results for the first paper

```
<FUNDING>
        <ORGANIZATION>Salus Mundi Foundation</ORGANIZATION>
        <ORGANIZATION>University of California</ORGANIZATION>
        <LOCATION>Berkeley</LOCATION>
        <ORGANIZATION>University of California</ORGANIZATION>
        <LOCATION>San Diego</LOCATION>
        <PERSON>Daniel C. Hallin</PERSON>
        <ORGANIZATION>Center for Advanced Study in the
        Behavioral Sciences</ORGANIZATION>
        <ORGANIZATION>Stanford University</ORGANIZATION>
</FUNDING>
```

Code Snippet 5: Results of the RegEx implementation for the first paper

```
<FUNDING>
        <ENTITY>Salus Mundi Foundation and</ENTITY>
        <ENTITY>University of California</ENTITY>
        <ENTITY>Berkeley and University of California</ENTITY>
        <ENTITY>San Diego</ENTITY>
        <ENTITY>Daniel C. Hallin </ENTITY>
        <ENTITY>Center for Advanced Study in the Behavioral
        Sciences</ENTITY>
        <ENTITY>Stanford University</ENTITY>
</FUNDING>
```

Code Snippet 6: Results of the NER implementation for the first paper

```
<FUNDING>
        <ORGANIZATION>Salus Mundi Foundation</ORGANIZATION>
        <ORGANIZATION>University of California</ORGANIZATION>
        <ORGANIZATION>Berkeley and University of California
        </ORGANIZATION>
        <LOCATION>San Diego</LOCATION>
        <PERSON>Daniel C. Hallin</PERSON>
        <ORGANIZATION>Fellow</ORGANIZATION>
        <ORGANIZATION>Center for Advanced Study
        </ORGANIZATION>
        <ORGANIZATION>Behavioral Sciences at
        Stanford University</ORGANIZATION>
</FUNDING>
```

Enumerations are hard to solve especially the last two elements that are connected with"and"/"or" are not split correctly. NER could probably learn this with more training in different cases. The last, very long entity name is correctly recognized by the RegEx approach. Here the NER approach fails. The NER approach does not have an entity with the"in the" conjunction in its training set. This leads to the model not recognizing the whole entity as one but splitting it up into two parts. It also does not recognize Stanford University as an extra entity. For this specific paper the RegEx approach delivers more accurate results than the NER approach. But the RegEx was explicitly built with this paper's funding text structure in mind.

6.1.2 Funding Section with Special Characters

The second paper by G. Sarchielli et al. does not only contain many different special characters, but also entity names in a different language. Although the paper is written in English, large portions of the funding section are in Italian and could be an issue. The funding text reads as follows:

"The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study has been financially supported by the Research Program of Region-University 2010-2012, Area 2-Ricerca per il Governo Clinico-Regione Emilia-Romagna. Project title: "Stili di direzione e di gestione delle risorse umane dipartimentali" promoted by the University Hospital St. Orsola-Malpighi Polyclinic and directed by Guido Sarchielli (Department of Psychology, University of Bologna)"[28]

Based on this text, the golden standard data was created manually (Code Snippet 7). All entities that can be categorized into one of the three defined categories are contained in this data.

Code Snippet 7: Gold standard result for the second paper

```
<FUNDING>
        <ORGANIZATION>Research Program of Region-University
        2010-2012</ORGANIZATION>
        <ORGANIZATION>Area 2-Ricerca per il Governo Clinico-Regione
        Emilia-Romagna</ORGANIZATION>
        <ORGANIZATION>University Hospital St. Orsola-Malpighi
        Polyclinic</ORGANIZATION>
        <PERSON>Guido Sarchielli</PERSON>
        <ORGANIZATION>Department of Psychology</ORGANIZATION>
        <ORGANIZATION>University of Bologna</ORGANIZATION>
</FUNDING>
```

The results of the RegEx approach contain at least parts of the relevant data. As seen in 8 the first match is missing the years. This is because the RegEx in its current form does not match numbers in the entity names. The second entity is matched correctly because it has a simple structure the RegEx can match. Matches three and four should be combined to form the gold standard. However because the RegEx does not recognize "St." as a conjunction word it can not match this entity correctly. In contrast the last result is a combination of multiple entities. Because the RegEx is built to match everything in trailing parentheses after a recognized entity.

Code Snippet 8: Results of the untweaked RegEx implementation

```
<FUNDING>
        <ENTITY>Research Program of Region-University</ENTITY>
        <ENTITY>Governo Clinico-Regione Emilia-Romagna</ENTITY>
        <ENTITY>University Hospital St</ENTITY>
        <ENTITY>Orsola-Malpighi Polyclinic and</ENTITY>
        <ENTITY>Guido Sarchielli (Department of Psychology,
        University of Bologna)</ENTITY>
</FUNDING>
```

With the NER approach the categorized results (Code Snippet 9) look very different. The algorithm recognizes the first two entities correctly. The third result is not in the golden standard, because it does not match one of the three defined categories. It would be necessary to define a new category and train the model accordingly. Result four is again an exact match with the golden standard. But the last two extracted entities are not recognized correctly. The model apparently has not learned or seen this particular structure before and does therefore not extract the three entities separately.

Code Snippet 9: Results of the untrained NER implementation

```
<FUNDING>
        <ORGANIZATION>Research Program of Region-University
        2010-2012</ORGANIZATION>
        <ORGANIZATION>Area 2-Ricerca per il Governo Clinico-Regione
        Emilia-Romagna</ORGANIZATION>
        <ORGANIZATION>University Hospital St. Orsola-Malpighi
        Polyclinic</ORGANIZATION>
        <PERSON>Guido Sarchielli</PERSON>
        <ORGANIZATION>Department of Psychology</ORGANIZATION>
        <ORGANIZATION>University of Bologna</ORGANIZATION>
</FUNDING>
```

Because both approaches produce inaccurate results in certain cases, the used RegEx was tweaked and the NER model was trained with the gold standard. The RegEx approach can now recognize "St." as a conjunction and can now match numbers in the entity name. The NER model was trained with the gold standard and is now able to extract the entities with a 100% accuracy. This also means that it no longer extracts the entity "Project title: "Stili di direzi-one e di gestione delle risorse umane dipartimentali" promoted". If a new category for project names would be introduced the model could be trained to recognize this entity, too.

With little tweaking the RegEx approach now produces more exact results (Code Snippet 10). By training the NER model with this new structure it can use the learned information for extraction on other papers which increases the accuracy of the results (Code Snippet 11). Both results are now much closer to the gold standard. The NER approach even coincides completely with the gold standard results. With more tweaks to how the RegEx approach handles parentheses and their content.

Code Snippet 10: Results of the tweaked RegEx implementation

```
<FUNDING>
        <ENTITY>Research Program of Region-University</ENTITY>
        <ENTITY>Governo Clinico-Regione Emilia-Romagna</ENTITY>
        <ENTITY>University Hospital St. Orsola-Malpighi
        Polyclinic and </ENTITY>
        <ENTITY>Guido Sarchielli (Department of Psychology,
        University of Bologna)</ENTITY>
</FUNDING>
```

Code Snippet 11: Results of the trained NER implementation

```
<FUNDING>
        <ORGANIZATION>Research Program of Region-University
        2010-2012</ORGANIZATION>
        <ORGANIZATION>Area 2-Ricerca per il Governo Clinico-Regione
        Emilia-Romagna</ORGANIZATION>
        <ORGANIZATION>University Hospital St. Orsola-Malpighi
        Polyclinic</ORGANIZATION>
        <PERSON>Guido Sarchielli</PERSON>
        <ORGANIZATION>Department of Psychology</ORGANIZATION>
        <ORGANIZATION>University of Bologna</ORGANIZATION>
</FUNDING>
```

6.2 Accuracy of the Approaches

To compare the RegEx and NER implementations their accuracy was measured. A gold standard was defined and the results of the approaches were matched with it. In Figure 10 these results are shown. The results are the average of the accuracy of four entity extraction on different papers [27], [28], [29], [30]. This was done without the tweaks and training shown in section 6.1.2.

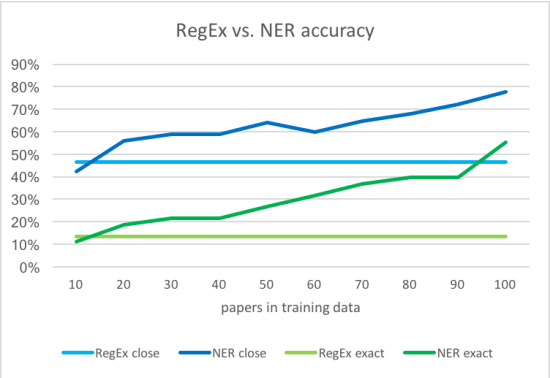

Figure 10: Accuracy comparison between RegEx and NER

The blue lines show close matches to the perfect output. Close matches are results that are only a partial match with a perfect output and are missing some part of the full output. The green lines show the exact matches. It is important to note that NER produces results with tags, which are also checked for correctness, this cannot be done and is not checked for the RegEx results. This means that if a result is tagged with the wrong category it does not count as a close or exact match for the NER approach because a wrongly categorized result is a wrong result, even if the entity name was extracted correctly.

As Figure 10 shows NER performs constantly better than RegEx. These numbers are averages from four measurements with different papers. For some papers the RegEx does not perform well at all with 0% exact matches and only 20% close matches. This means it only recognized two out of five entities correctly for those particular papers. In comparison NER with the fully trained model achieves 90% close matches and 80% exact matches for the same paper. Figure 10 also shows that NER can learn to recognize the basic structures of funding text quickly with very little training data. The big jump from 90 to 100 training data papers stems from a specific structure used in the last 10 papers that matches the structure in one of the tested papers and therefore NER learned how to recognize the entities in exactly these papers.

The performance of the RegEx implementation stays constant because it is not able to learn from its own results. The RegEx has to be manually curated in order to constantly work properly, whereas with enough training data the NER can learn by itself. If it does need help from a developer it is much easier to adjust the training data and model.

It is important to note that Giles & Councill say they achieved an accuracy of >90% with a regular expression [16]. This could mean that they measured their accuracy differently or that this expression is not fully optimized and should be tweaked further if used productively.

6.3 Discussion

To compare the developed approaches and the earlier formulated goal the four properties introduced to evaluate the existing approaches are used again. Figure 11 shows the goal drawn over the properties of the NER and RegEx implementations.

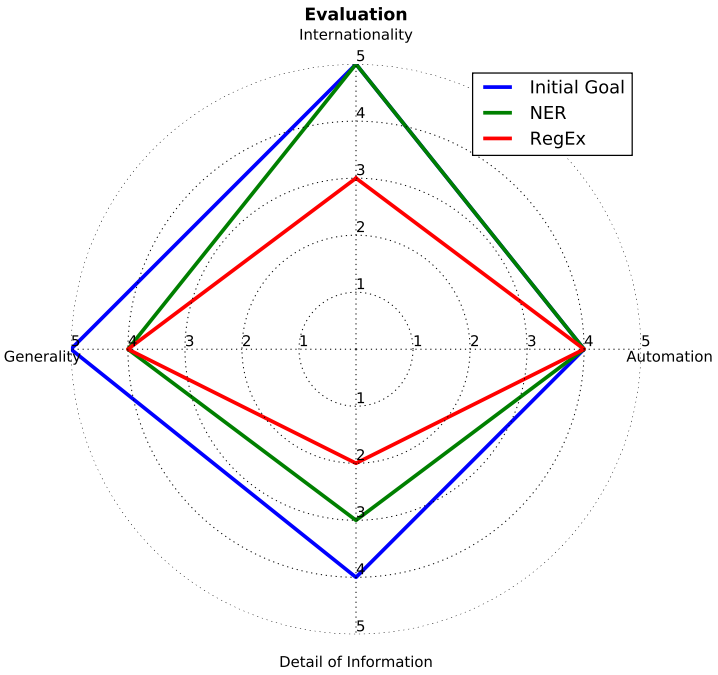

Figure 11: Properties of the developed approach(es) compared to the initital goal

In terms of internationality the NER implementation reaches the goal. It can be adapted to other languages by changing the regular expression used to find the funding section as well as training a new model to recognize funding entities in the new language. Adapting the RegEx implementation to a new language is harder. Many languages do not have the same capitalization rules this implementation uses to recognize the entities in English. Depending on the new language it may be impossible to differentiate a funding entity from the rest of a sentence without looking at the context. Due to this the RegEx implementation is rated lower in this property and does not meet the goal.

Funding and acknowledgement sections in papers from different scientific disciplines are very similar in structure. The RegEx implementation as well as the NER implementation can extract funding entities from scientific papers independent of discipline. Both approaches however cannot analyse papers in which the funding section is not explicitly declared with a title. This is because the initial section extraction of the approaches, in which the funding section is found, uses a regular expression that matches the section title. If there is no section title, it cannot find a match and thus not extract the relevant section.

Both approaches meet the set goal for the automation property. The process of reading and analysing a PDF is fully automated. A user only has to provide a valid PDF and both approaches will return their results in form of an XML file.

What is not automated is the gathering and downloading of the PDFs and the data is not validated or cross-referenced with other sources. Based on this both approaches are rated with a 4 in the automation property.

The level of detail of the extracted data is the biggest difference between the two approaches. Because the RegEx implementation only extracts the funding entities from the text and cannot add any context to the results, the level of detail is minimal and limited to the names of the entities. The NER implementation is able to categorize the results. With categorized results it is easier to use the results for further computations compared to the results of the RegEx implementation. Even though the NER implementation performs better than its competitor in this property it does not reach the goal for this property. It is able to extract most of the data from the paper but the level of detail could still be increased by cross-referencing the extracted data with databases from funders and publishers.

Overall the NER implementation outperforms the RegEx implementation and is easier to train for new situations. Especially with a bigger training set for the machine learning algorithm the RegEx implementation will not be able to compete. In this state however, both approaches still have issues that need to be fixed in order to get an approach that meets all the properties of the initial goal.

7 Conclusion

Two methods were developed that can both extract funding entities from papers. One method is using RegEx to recognize entities in the funding section of a paper. The other method uses a machine learning algorithm for NER developed by the Stanford NLP team. The RegEx implementation delivers accurate results if the funding text is well structured and contains clearly separated entity names. To extract entities from more complex structures the developed RegEx has to be tweaked, so it can match the patterns of the structure. This happens directly in the program code. In comparison the NER implementation delivers similarly accurate results for simple structures it has learned before. If it encounters a new structure that was not included in the training set, the results become inconsistent. In order to be able to extract entity names from structures that were not trained, the training set can be extended. This way it is possible to train the model on the new structure without changing the program code. This improves the method's accuracy for the new and similar structures.

Results of the extraction are saved to an XML file. The XML format allows for further computation of the results. Every entity the RegEx method extracts, is saved with an "Entity" tag. The NER implementation is able to categorize its results, which adds more information and allows for better handling of the data.

Both methods fully automate the funding entity extraction from a PDF file. However, not all PDFs can be read. This hurts the general applicability of the two approaches. Even though both approaches do not meet the goal for the level of detail that was set, the results produced by the NER method contain more detail. This and the fact that it can be trained easily make the developed NER approach better than the RegEx method.

The developed approaches lay the foundation to make the funding in science more transparent and thus could strengthen the trust the people have in science.

8 Outlook

The problem of transparency in science funding can not be solved with this approach alone. Information on funding for scientific research needs to be declared explicitly and, in the best case, made easily accessible for the public in full detail. With full funding transparency the trust in scientific results can be improved and publishing biased papers becomes harder. This could possibly make it less attractive for companies to influence the results.

The data extraction methods developed in this thesis are a first step to bring more transparency in to the funding landscape of science. In a next step it is important to further develop the approach to achieve a higher level of detail in the results. The results can be used in various ways, they could be combined with other metadata about the paper in order to be able to analyse funding in combination with authors, keywords or even countries.

In order to get more details and a higher accuracy for the NER algorithm a bigger training data set that potentially contains more tags needs to be generated. The results of the extraction could be used to get further information from online databases be it through available APIs or web scraping. There are not only the

databases from publishers and grants, but also harder to find databases focusing on papers from specific disciplines or universities that could be integrated in the approach. By cross-referencing the initial results with various databases a full picture of the funding of a paper can be achieved. This data could be used for various things.

8.1 Processing the Data

The data could be saved into a database and mined to maximize the gained information form the data. Wang & Shapira's research [8] could be expanded to be used not only for Nanotechnology papers but for scientific papers in general. The data could be analysed for patterns in the data, for example there could be certain funding entities only investing in papers with very specific keywords. This could help to identify big investors in different areas and make the funding in some disciplines more transparent.

The data could also be used to build maps similarly to what Boyack & Börner [17] did. A map or similar visualization of the connections between authors and funding entities could help to identify big players in the industry or upcoming trends and fields. This could help to predict the future hot topics in science and show where funding is needed the most. With good visualization the data could be used in many different ways to make scientific research more transparent.

8.2 Webservice

8.2.1 Searchable Database

In order to make the data more accessible for the public a service similar to CiteSeerX could be built around this approach. What CiteseerX does is extract metadata from papers and make it searchable through a web interface. With such a service the public could easily access funding information on authors, keywords or even projects. This would need a system that can extract the basic data from a PDF and then expand this data by pulling more information from grant, publisher and university websites and databases. The gathered data can then be written in a database and made available on the web service. Users could also be used to expand the database by allowing them to upload PDFs of scientific papers that could then be automatically analysed.

8.2.2 Bias Checker

A completely different (and more complex way to use this approach would be to use it as part of a 'bias checker'. A user could upload a PDF of a scientific paper and the service would extract the funding entities. Then the conclusion or result of the paper is analysed and evaluated. Based on the funding data and the result of the paper a bias evaluation is created. The user then gets a feedback on if the funding could have influenced the result of the paper. How the result of a paper can be analysed would be a problem of computer linguistics and is not easy to solve. Such a tool would help to make it less attractive for companies to influence the results of scientific research and could strengthen the trust in science.

Acknowledgments

I would like to acknowledge my familiy for the support while writing my bachelors thesis. I would also like to thank Manuel Keller and Gian Tschudi for proof reading. Special thanks go to Prof. Dr. Lorenz Hilty and Dr. Achim Schneider for supervising this work.

Conflicts of Interest

The author declares no potential conflicts of interest with respect to the research and/or authorship of this article.

Funding

This thesis emerged from research supported by the University of Zürich. It grew out of discussions with Prof. Dr. Lorenz M. Hilty from the Informatics and Sustainability Research Group at the University of Zürich and Achim Schneider, PhD, an external research advisor.

References

[1] Anglo-American Cataloguing Rules. (2016). AACR2, [Online]. Available: http://www.aacr2.org/about.html (visited on 08/12/2016).

[2] Library of Congress. (2016). MARC 21 format for bibliographic data: Table of contents (network development and MARC standards office, library of congress), [Online]. Available: http://www.loc.gov/marc/bibliographic/ (visited on 08/12/2016).

[3] Dublin Core. (2016). DCMI metadata basics, [Online]. Available: http://dublincore.org/metadata-basics/ (visited on 08/15/2016).

[4] Dublin Core. (2016). Dublin core metadata element set, version 1.1, [Online]. Available: http://dublincore.org/documents/dces/ (visited on 08/15/2016).

[5] T. B. Darmoni S. J. (2016). The use of dublin core metadata in a structured health resource guide on the internet, [Online]. Available: http://www.chu-rouen.fr/cismef/cismefdc.html (visited on 08/12/2016).

[6] SAGE Publications. (2016). Funding acknowledgements, SAGE publications, [Online]. Available: https://uk.sagepub.com/en-gb/eur/funding-acknowledgements (visited on 05/27/2016).

[7] Springer. (2016). Before you start, [Online]. Available: https://www.springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/before-you-start (visited on 05/27/2016).

[8] J. Wang and P. Shapira, "Funding acknowledgement analysis: An enhanced tool to investigate research sponsorship impacts: The case of nanotechnology", *Scientometrics*, vol. 87, no. 3, pp. 563–586, 2011. [Online]. Available: http://link.springer.com/article/10.1007/s11192-011-0362-5 (visited on 06/25/2016).

[9] SAGE Publications. (2016). SAGE publications ltd, [Online]. Available: https://uk.sagepub.com/en-gb/eur/home (visited on 07/13/2016).

[10] National Science Fund. (2016). NSF award search: Award#1346854, [Online]. Available: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1346854&HistoricalAwards=false (visited on 07/13/2016).

[11] Thomson Reuters. (2016). Web of science core collection, [Online]. Available: https://apps.webofknowledge.com (visited on 11/19/2016).

[12] US Government. (2016). Search grants, [Online]. Available: http://www.grants.gov/web/grants/search-grants.html (visited on 08/12/2016).

[13] S. Data. (2016). Uni-transparenz, [Online]. Available: http://srfdata.github.io/2016-04-uni-transparenz/ (visited on 05/29/2016).

[14] J. Wang and P. Shapira, "Is there a relationship between research sponsorship and publication impact? an analysis of funding acknowledgments in nanotechnology papers", *PLOS ONE*, vol. 10, no. 2, e0117727, Feb. 19, 2015, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0117727. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117727 (visited on 05/27/2016).

[15] T. G. u. J. Schmidli. (Apr. 18, 2016). Uni-bindungen: Was die daten aussagen können und was nicht, Schweizer Radio und Fernsehen (SRF), [Online]. Available: http://www.srf.ch/news/schweiz/uni-transparenz/uni-bindungen-was-die-daten-aussagen-koennen-und-was-nicht (visited on 08/12/2016).

[16] C. L. Giles and I. G. Councill, "Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing", *Proceedings of the national academy of sciences of the united states of america*, vol. 101, no. 51, pp. 17 599–17 604, 2004. [Online]. Available: http://www.pnas.org/content/101/51/17599.short (visited on 07/20/2016).

[17] K. W. Boyack and K. Börner, "Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers", *Journal of the american society for information science and technology*, vol. 54, no. 5, pp. 447–461, 2003. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/asi.10230/full (visited on 07/20/2016).

[18] U.S. National Library of Medicine. (2016). Computer retrieval of information on scientific projects source information (CRISP), Computer Retrieval of Information on Scientific Projects Source Information, [Online]. Available: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CSP/ (visited on 11/19/2016).

[19] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification", *Lingvisticae investigationes*, vol. 30, no. 1, pp. 3–26, 2007. [Online]. Available: http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002 (visited on 11/22/2016).

[20] Stanford University. (2016). The stanford natural language processing group, [Online]. Available: http://nlp.stanford.edu/software/CRF-NER.shtml (visited on 11/19/2016).

[21] P. Lopez. (2016). GROBID-NER, GitHub, [Online]. Available: https://github.com/kermitt2/grobid-ner (visited on 11/19/2016).

[22] P. Lopez. (2016). GROBID NER documentation, [Online]. Available: https://grobid-ner.readthedocs.io/en/latest/ (visited on 11/19/2016).

[23] P. Lopez. (2016). GROBID, GitHub, [Online]. Available: https://github.com/kermitt2/grobid (visited on 11/19/2016).

[24] Stanford University. (2016). The stanford natural language processing group, [Online]. Available: http://nlp.stanford.edu/software/CRF-NER.shtml#Models (visited on 11/19/2016).

[25] Text Encoding Initiative. (2016). TEI: Text encoding initiative, [Online]. Available: http://www.tei-c.org/index.xml (visited on 11/19/2016).

[26] Oracle. (2016). Pattern (java platform SE 7), [Online]. Available: https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html (visited on 11/22/2016).

[27] D. C. Hallin and C. L. Briggs, "Transcending the medical/media opposition in research on news coverage of health and medicine", *Media, culture & society*, vol. 37, no. 1, pp. 85–100, 2015. [Online]. Available: http://mcs.sagepub.com/content/37/1/85.short (visited on 11/19/2016).

[28] G. Sarchielli, G. De Plato, M. Cavalli, S. Albertini, I. Nonni, L. Bencivenni, A. Montali, A. Ventura, and F. Montali, "Is medical perspective on clinical governance practices associated with clinical unitsa performance and mortality? a cross-sectional study through a record-linkage procedure", *SAGE open medicine*, vol. 4, p. 2 050 312 116 660 115, 2016. [Online]. Available: http://smo.%20pub.com/content/4/2050312116660115.full (visited on 11/22/2016).

[29] S. M. Somerset, "Refined sugar intake in australian children", *Public health nutrition*, vol. 6, no. 8, Dec. 2003, ISSN: 1368-9800, 1475-2727. DOI: 10.1079/PHN2003501. [Online]. Available: http://www.journals.cambridge.org/abstract_S1368980003001083 (visited on 11/26/2016).

[30] D. A. Lawson, N. R. Bhakta, K. Kessenbrock, K. D. Prummel, Y. Yu, K. Takai, A. Zhou, H. Eyob, S. Balakrishnan, C.-Y. Wang, P. Yaswen, A. Goga, and Z. Werb, "Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells", *Nature*, vol. 526, no. 7571, pp. 131–135, Sep. 23, 2015, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature15260. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature15260 (visited on 11/26/2016).

Appendix

Prototypes

The source code for the prototypes can be found on the enclosed CD or on GitHub[13]. The protoypes also contain all the used models and gold standard files. There are no scientific papers included, these have to be downloaded manually. In order to start an analysis, the PDFs can be put into a folder called "input" and the results will be written to a folder called "output" in the prototype's root directory. If these folders do not exist they have to be created manually. The project is based on Maven however, not all used packages are available. Especially Stanford NER has to be downloaded separately from the Stanford NLP website[14]. This website also contains instructions how to use the NER application to generate training data.

The used packages to extract text from PDFs are PDFBox[15] and PDFxStream[16]. PDFxStream has a limitation to the amount of PDFs it can extract at a time. If it were used for bigger scale projects, the full version would have to be bought or an alternative used.

Training Data

Also included on the enclosed CD are references to the 100 papers that were used to create the training data as well as the training data itself. The training data file currently has close to 5400 lines and can be extended with more data if needed.

Radar graphs

To draw the radar graphs shown in this paper a small python application was used. It can also be found on the CD. It uses matplotlib[17] to draw the graphs based on input data from an XML file. An example XML file is included in the source code.

Accuracy raw Data

The raw data used to generate Figure 10 can be found here:

[13]https://github.com/dikohl/funding-extractor
[14]http://nlp.stanford.edu/software/CRF-NER.shtml
[15]https://pdfbox.apache.org/
[16]https://www.snowtide.com/
[17]http://matplotlib.org/

Hallin & Briggs

training size	RegEx close	NER close	RegEx exact	NER exact
10	87.5%	37.5%	37.5%	12.5%
20	87.5%	50%	37.5%	25%
30	87.5%	62.5%	37.5%	37.5%
40	87.5%	62.5%	37.5%	37.5%
50	87.5%	62.5%	37.5%	37.5%
60	87.5%	62.5%	37.5%	37.5%
70	87.5%	62.5%	37.5%	37.5%
80	87.5%	62.5%	37.5%	37.5%
90	87.5%	62.5%	37.5%	37.5%
100	87.5%	75%	37.5%	50%

Sarchielli et al.

training size	RegEx close	NER close	RegEx exact	NER exact
10	12.5%	25%	0%	12.5%
20	12.5%	50%	0%	12.5%
30	12.5%	50%	0%	12.5%
40	12.5%	50%	0%	12.5%
50	12.5%	50%	0%	12.5%
60	12.5%	50%	0%	12.5%
70	12.5%	50%	0%	12.5%
80	12.5%	62.5%	0%	25%
90	12.5%	62.5%	0%	25%
100	12.5%	62.5%	0%	25%

S. M. Somerset

training size	RegEx close	NER close	RegEx exact	NER exact
10	66.66%	66.66%	16.66%	0%
20	66.66%	83.33%	16.66%	16.66%
30	66.66%	83.33%	16.66%	16.66%
40	66.66%	83.33%	16.66%	16.66%
50	66.66%	83.33%	16.66%	16.66%
60	66.66%	66.66%	16.66%	16.66%
70	66.66%	66.66%	16.66%	16.66%
80	66.66%	66.66%	16.66%	16.66%
90	66.66%	83.33%	16.66%	16.66%
100	66.66%	83.33%	16.66%	66.66%

Lawson et al.

training size	RegEx close	NER close	RegEx exact	NER exact
10	20%	40%	0%	20%
20	20%	40%	0%	20%
30	20%	40%	0%	20%
40	20%	40%	0%	20%
50	20%	60%	0%	40%
60	20%	60%	0%	60%
70	20%	80%	0%	80%
80	20%	80%	0%	80%
90	20%	80%	0%	80%
100	20%	90%	0%	80%