Universiteit Leiden

Master Computer Science

Data mining paper meta-data: a novel model for large scale paper meta-data research

Name:	Patrick Bergman BSc, BBA
Student ID:	s1553097
Date:	[March 20, 2022]
Specialisation:	Advanced Data Analytics
1st supervisor:	prof.dr. M.E.H. van Reisen
2nd supervisor:	dr. E.M. van Mulligen (Erasmus MC)

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgements

First of all I want to thank Mirjam van Reisen for sticking with me and giving me the time to finish this thesis. She pushed where necessary and helped where appropriate. Furthermore, I want to thank Erik van Mulligen for tips, directions, and help with my process.

Finally, I want to thank my mother for her everlasting belief in me. Unfortunately she passed away during this thesis research, but was always with me in my thoughts.

Abstract

Doing large scale research on meta-data of published papers is cumbersome. In general the larger databases containing scientific papers are well known. Examples are Google Scholar, Scopus and JSTOR. Unfortunatly Microsoft acadamic retired at the 31st of december 2021. From these examples only Microsoft Academic has a seperate API to retrieve data. But only the data they were able to scrape from the sources. Dedicated meta-data that describe papers are scarce. There are services available that provide meta-data, like Crossref. But you will need DOI's of the specific papers you want to investigate.

This thesis describes a structured model to get meta-data in order to make finding and analysing meta-data easier for future researches. Through several steps meta-data are retrieved, processed and made available to be analysed. This model will be focused on analysing geographical data. A use case will show how the model can be used in a research. This use case will research the usage of FAIR in papers by reviewing the geographical location of the first authors' affiliation. The resulting data is grouped per year since the publication of the original FAIR paper and will display the geographical usage.
A sub-question will research the publishers and the amount of meta-data that they share.

The model, that contains six steps, and the use case will demonstrate that it is no easy task to do research on the meta-data of large amounts of papers. An earlier published paper did a similar research, but by hand and with a maximum of 100 papers. This model can be of help for researchers that intent to research large amounts of papers which can contain hundreds of papers, but also shows the limitations of doing such a research.

In the end this thesis will also show how the agenda setting model, developed by Kingdon in 1984, is still relevant for issues we face today. There is a world-wide problem of researchers that are either doing research that already has been done, because they were not able to find it or researchers that cannot find data from previous researches to accelerate their own research. The FAIR framework offers convenient policies for research to counter this problem. Politically this is recognised by the European Union which requires the use of the FAIR framework for certain grants. Combined this can develop into an agenda where the FAIR framework and sharing data can become mandatory while accelerating research by collaborating.

Contents

Chapter 1

Introduction

There is a lot of information available in the world. Most of this information is gathered on the world wide internet and is expanding each day. It is estimated that by the year 2025 over 200 zettabyte of data will be on the internet[1]. But only a part of this is scientific data shared by institutions, researchers and companies. In order to do a literature study researchers must know where to look and what they are looking for.

In 2020 van Reisen et al. published a paper that researches the geographical spreading of FAIR implementation [Rei+20]. The conclusion of their research is 'The FAIR Guiding Principles have experienced significant expansion in acceptance and implementation, although implementation is largely limited to the Western hemisphere and to bio- and natural sciences (95% of articles reviewed)'[Rei+20]. While this is a good conclusion of their research, they only used 100 articles as a reference. They manually processed each article on it's own by hand, which is a time consuming accomplishment.
In their research they used a naive approach: 'To investigate FAIR implementation, a literature review of 100 randomly selected academic journal articles – citing the founding article, was conducted'[Rei+20]. This is also a necessity since google scholar does not have an API with available meta-data. Other services, like Crossref, do have meta-data on published articles. The downside is that you have to learn how to use another database to find you are looking for. Especially since Crossref required using DOI's (Digital Object Identifiers) to find meta-data and Google scholar does not provide DOI's in their results.

Such a research approach shows how difficult it is to do a thorough literature research. As far as we know, there is no method available to do a cross-database literature research on a certain topic. This results in the following main research question:

"How can one do a thorough, literature research across different databases and data-sets?"
For this research Google Scholar, Crossref, ScraperAPI, PHP and Python are used with webmining techniques. In order to answer the main research question a case study which encounters the problem as described above and will provide an answer to the main research question. The research question for the case study is:

"What is the geographical spread of FAIR based on papers that cite the original FAIR paper?"

This case is similar compared to the aforementioned paper by van Reisen et al. The difference is the systematic and programmatic approach and this case study is just a supportive study in order to answer the main research question.

[1] https://cybersecurityventures.com/the-world-will-store-200-zettabytes-of-data-by-2025/

Chapter 2

Problem statement

FAIR, which is an acronym for findability, accessibility, interoperability, and reuse of digital assets, was introduced in the paper 'FAIR Guiding Principles for Scientific Data Management and Stewardship', published by *Scientific Data*. *Scientific Data* is a magazine which is part of the journal Nature. In the original FAIR paper[Wil+16], it is suggested that data as a research object must adhere to certain standards. Specifically, the authors stress the need to make scientific data used in research reusable for machines and interpretable by humans.
FAIR intents to make acquired data (often acquired with great effort and/or cost) available for later use. This preservation of data will be done by the best possible standards to maintain the valuable assets.

FAIR principles are used by the University of Leiden, where this thesis has been researched, mainly by the Computer Science department. An article about their importance and usage is published on the website under the title 'Leiden: Silicon Valley of FAIR data'[2]. Nowadays there is also a larger international initiative based in Europe called GO-FAIR. This 'offers an open and inclusive ecosystem for individuals, institutions and organisations working together through implementation networks'.

Sharing data between researchers is important in order to avoid double researches and acquiring the same data in multiple researches. A method to gather data for a literature study is not known at the time of this thesis. In a previous research by van Reissen et al.[Rei+20] in 2020 a manual method is used by randomly selecting papers from Google scholar and anaylsing them by hand.

[2] https://www.universiteitleiden.nl/en/research-dossiers/data-science/leiden-silicon-valley-of-fairdata

Chapter 3

Theory

This chapter will briefly describe the FAIR framework and why it is important for researchers to become familiarized with it. This is important to show the relevance of the guiding principles provided by FAIR. Next the concept of web mining is explained, along with why this technique has been used for this research. Finally, Kingdon's agenda-setting model is described and put into the context of this research, which involves reviewing the results from the mining process in combination with the results.

3.1 The FAIR framework

Each year, more and more data is collected by companies, governments and researchers. This data is created or discovered in numerous ways, saved in different locations, and then either distributed or not to different places in the digital world. All this data can contain potential knowledge and interesting patterns. By using a shared framework, such as the FAIR principles, data can be found, processed, and shared according to a common approach. Every researcher from any field could benefit from a uniform method of sharing data, helping to improve their research and achieve better results.

As mentioned earlier, the acronym FAIR stands for 'findable, accessible, interoperable and reusable': qualities that enable the use and reuse of data by machines and persons [Wil+16]. The authors of the term identified certain obstacles arising from the fact that 'science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments'[Wil+16]. 'Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data'[Wil+16].

According to Wilkinson et al. [Wil+16], the FAIR framework or principles 'serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added value gained by contemporary, formal scholarly digital publishing'. It should be noted that FAIR is not intended to be a standard, but more a collection of principles that can improve the re-use of data[Mon18].

In summary, the FAIR framework provides a collection of principles that can aid scientific data management and stewardship in the digital age we now live in. The framework can be applied to all kinds of different research libraries and information resources to improve their data research and share scientific data with others.

3.2 Web content and structure mining

When a researcher applies data mining techniques to the content and structure of the web, it is referred to as web mining. These techniques can help identify local and global structures via the internet. In the data mining field these are also referred to as ‘patterns’ or ‘models’ [Han07]. Web mining can profit from various types of data, both structured and unstructured, while in general data mining researchers only tend to deal with structured data. Web mining is therefore helpful in transforming human-understandable content into machine interpretable data.

There are three distinguishable types of web mining: content, structure and usage mining. Content mining, the first type, is a form of text mining. Individual web pages are the primary resource for content mining. A more detailed explanation of text mining on the web has been written by Chakrabarti et al.[Cha00]. The HTML of web pages are used not only for layout, but also denote a logical structure. This logical structure can be utilised to find specific types of data a researcher is looking for, for example images, texts, or hyperlinks.

Web structure mining is another technique used in this research. Usually such structure mining utilises the hyperlinks that are present on each web page. A great example is the PageRank algorithm which made Google so successful. This algorithm determines the importance of a link by the number of other hyperlinks that refer to the page from other web pages, particularly relevant ones [Pag+99].

The structure of a single web page can be analysed as well and provide data regarding its function; or, in the case of Google Scholar, a list of relevant papers according to your search. In a paper by Cooley, five types of web page are defined: ‘head’ pages, ‘navigation’ pages, ‘content’ pages, ‘look-up’ pages and ‘personal’ pages [CMS99]. In this research only ‘head’ pages, which generally serve as entry points for a specific website, and ‘navigation’ pages, are used. Navigation pages contain mainly hyperlinks to other web pages, which is used in usage mining.

3.3 Kingdon’s agenda setting model

John W. Kingdon describes in his book *Agendas, Alternatives, and Public Policies* [KS84] the process of problems or subjects that are scheduled by the government on their agendas. The stages before such topics are scheduled have a deciding influence on the sequel of the policy cycle and the final results. Kingdon assumes three different streams: the problem stream, the policy stream, and the political stream. These are independent from one another and develop separately towards the final policy agenda [KS84].

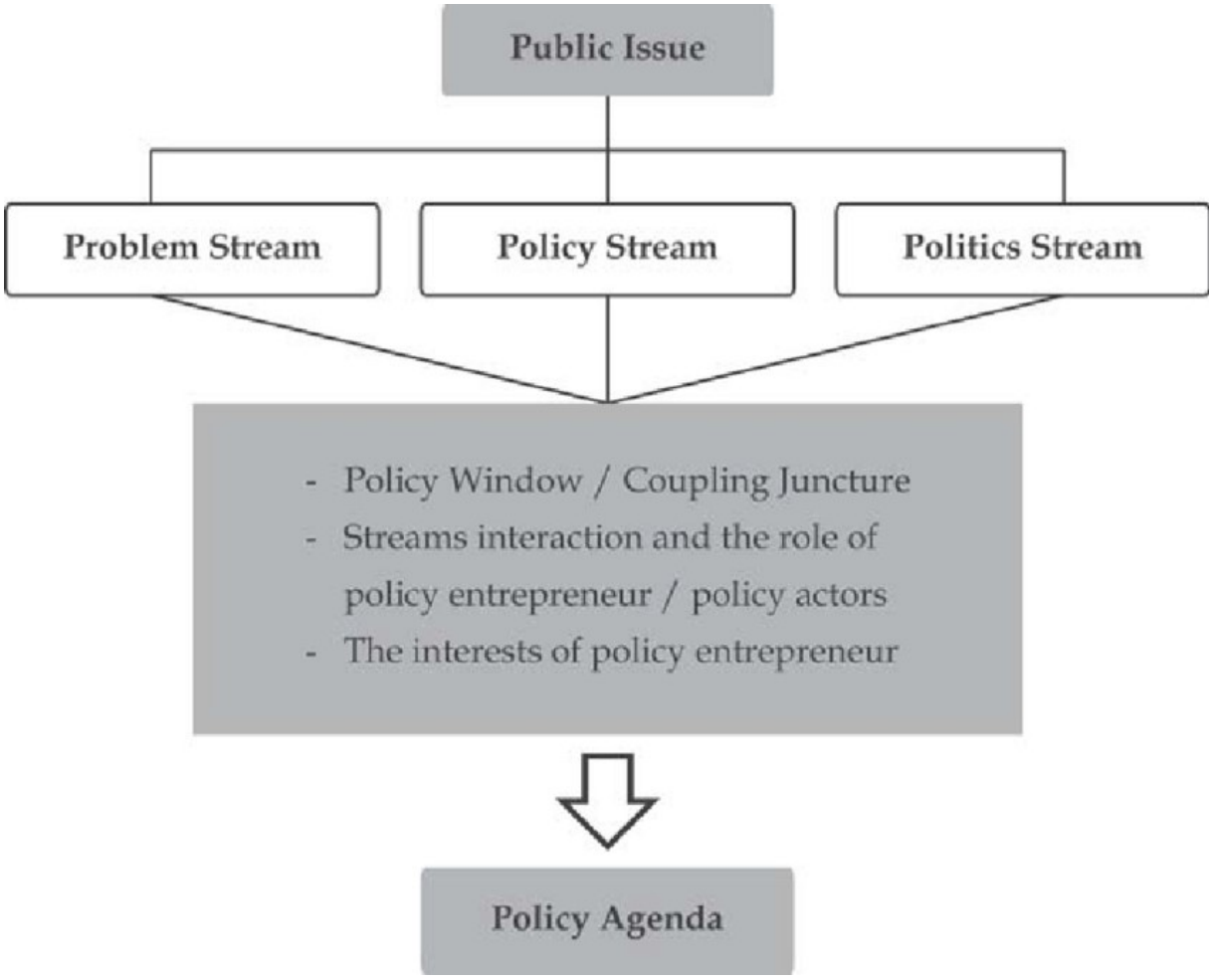


Figure 3.1: Kingdon's agenda setting model[Wid18]

3.3.1 Problem Stream

The left stream as displayed in Figure 3.1 of Kingdon's model is the problem stream. A problem is defined by Kingdon as follows: 'Conditions become defined as problems when we come to believe that we should do something about them' [Kin11]. A problem stream will be of use when governments and the people that work for them identify public issues as a problem and, as defined by Kingdon, decide they should do something about them. The time in which an issue can become a problem is often short; it should be linked with the other two streams so the results can be used to form a policy agenda.

3.3.2 Policy Stream

The central stream in the model as displayed in Figure 3.1 is the policy stream. In this stream governments and their employees conceive of policies which in turn are proposed. These conceptual policies can still be influenced by people outside of the government. Not only lobbyists, but also researchers and other scientific institutes may be consulted to form these policies. These policies are intended to solve the problems as defined in the former problem stream. To counter resistance or hesitation regarding the proposed policies, some

experimental partial policies/actions may sometimes be taken in order to soften the impact of the full policy and so people can get used to them [Kin11].

3.3.3 Politics Stream

The final stream is the politics stream. Kingdon divides this into three part, 'the national mood', 'the organised political forces', and 'changes within government'. These parts together basically represent how the wind blows, what people's feelings are towards the agenda, and whether there is willingness among the people and politicians to develop an agenda to solve the public issue at hand. The short period in which this is possible is denoted as a 'policy window' [Qui86]. Striving for consensus among all involved parties is the common goal in this stream [Kin11].

3.3.4 Policy Agenda

After the time is right and these three streams become joined, a window of opportunity will develop. When the policy agenda is formed, entrepreneurs should use this window of opportunity to manage their resources, such as time, reputation, money, and energy, to carry out the policy as formed by the government. These entrepreneurs may provide practical solutions to the first identified public issue. This window of opportunity does not stay open for a long time and an entrepreneur must act quickly when all these streams are joined [Kin11].

Chapter 4

Methods

This section describes the methods used to retrieve and analyse the data. First, it describes the usage of Google Scholar and how data is retrieved without the availability of an API. It then explains how DOIs are found with the help of the Google Scholar results. A subsection outlines how the Crossref this service is used and for what purpose. Finally, the methods for finding geolocations and utilizing the metadata are delineated.

4.1 Used programming language's

For this research two different programming languages are used, PHP and Python. Each has its strengths and weaknesses and should be used accordingly. PHP, designed in 1994 by Rasmus Ledorf, is a programming language specific for the web and is inspired by Perl (A program language developed in 1987). This language was chosen for this study due to its tight integration with web technologies and its suitability for retrieving internet data, which is the heart of this study.

Python is the other language of choice due to its tight integration with data analytics and the scientific world. Python is a well-documented and universally used language to process data and there are many packages available to assist in data analytics. In this case it is used to analyse and visualise the data that is retrieved from the internet.

4.2 Scraper API

To retrieve HTML from a webpage, a service is used called ScraperAPI. According to its producer, 'ScraperAPI handles proxies, browsers, and CAPTCHAs, so you can get the HTML from any web page with a simple API call!' [LLC21]. ScraperAPI takes care of the language on the request and thus always obtains results in English. They also claim automatic reCAPTCHA handling, which is requested by Google Scholar when you make more requests then their terms of service allow.

4.3 Using Google Scholar

Google Scholar was first introduced in November 2004. It made it easy to find academic information about a wide range of subjects, which was particularly important for those who did not have access to a broad spectrum of fee-based indexing/abstracting databases [Jac05]. At the time, these were the main routes used to discover new information.

Google Scholar is now a place where one can find a large collection of scientific papers with a mouse click. For this research, it was used to locate papers that cite the first FAIR proposal

paper by Wilkinson et al. [Wil+16]. This is possible thanks to the option to view all papers that cite a paper according to Google.

Unfortunately Google shows a maximum number of 10 pages, with 10 papers displayed on each. This results in a maximum availability of 100 papers. In Google Scholar it is possible to filter by year, however, and since the FAIR paper by Wilkinson et al. was introduced in 2016, this was done for each year from 2016 to 2021.

A short program was created in PHP to retrieve the HTML via Scraper API of each page of papers discovered that cite Wilkinson et al., filtered for each year since that paper was published. The resulting HTML was stored in the database for later use. Specifically, only the HTML for the first page was retrieved, after which the number of subsequent pages was determined, because there could also be fewer than 10 pages available. This 10 pages limit is set by Google for unknown reasons. For each of the subsequent pages, the URL was adjusted to retrieve the specific HTML for that year and that page.

4.4 Retrieving DOI's

A DOI is an abbreviation for Digital Object Identifier, which is commonly used in the scientific world to reference papers. DOIs are part of the DOI System, which 'is a managed system for persistent identification of content on digital networks. It can be used to identify physical, digital, or abstract entities' [Pas10]. The DOIs of papers can help find related metadata with Crossref, which will be explained in the next part of this chapter.

In order to retrieve the DOIs from each link related to a document that cites the original FAIR paper, a non-standard approach is necessary. For the 21 most common domains that were found in the Google Scholar HTML pages, a specific approach was developed to retrieve possible DOIs. This required the retrieval of the HTML of the specific domains. The results were then filtered with the help of a DOM-Crawler, specifically the Symfony DOM-Crawler[3]. This component contains functions to make filtering HTML easier.

The approach taken to retrieve other DOIs from sites that are not represented in these 21 domains was straightforward. The HTML from a specific url of a paper was retrieved with the ScraperAPI and analysed with the DOM-Crawler. This involves a search for a meta tag with the name of 'doi' which often contains the DOI of the corresponding paper. There are also cases where this HTML-tag is not present and will not be discovered.

4.5 CrossRef

In 2000 a non-profit organisation was founded that centralizes DOIs and their corresponding metadata. On their homepage they state that 'Crossref makes research outputs easy to find, cite, link, assess, and reuse'[4].

[3] https://symfony.com/doc/current/components/dom crawler.html
[4] https://www.crossref.org/

Publishers initiated the Crossref organization in order 'to provide a service that would enable publishers to link to each other persistently in the age of online publishing' [Lam15]. When the online space for publishing grew, the phenomenon of 'link rot' appeared. This basically meant that content would move quite often, resulting in links in papers no longer working [Lam15]. An idea of a basic workflow in which Crossref is used is displayed in Figure 4.1.

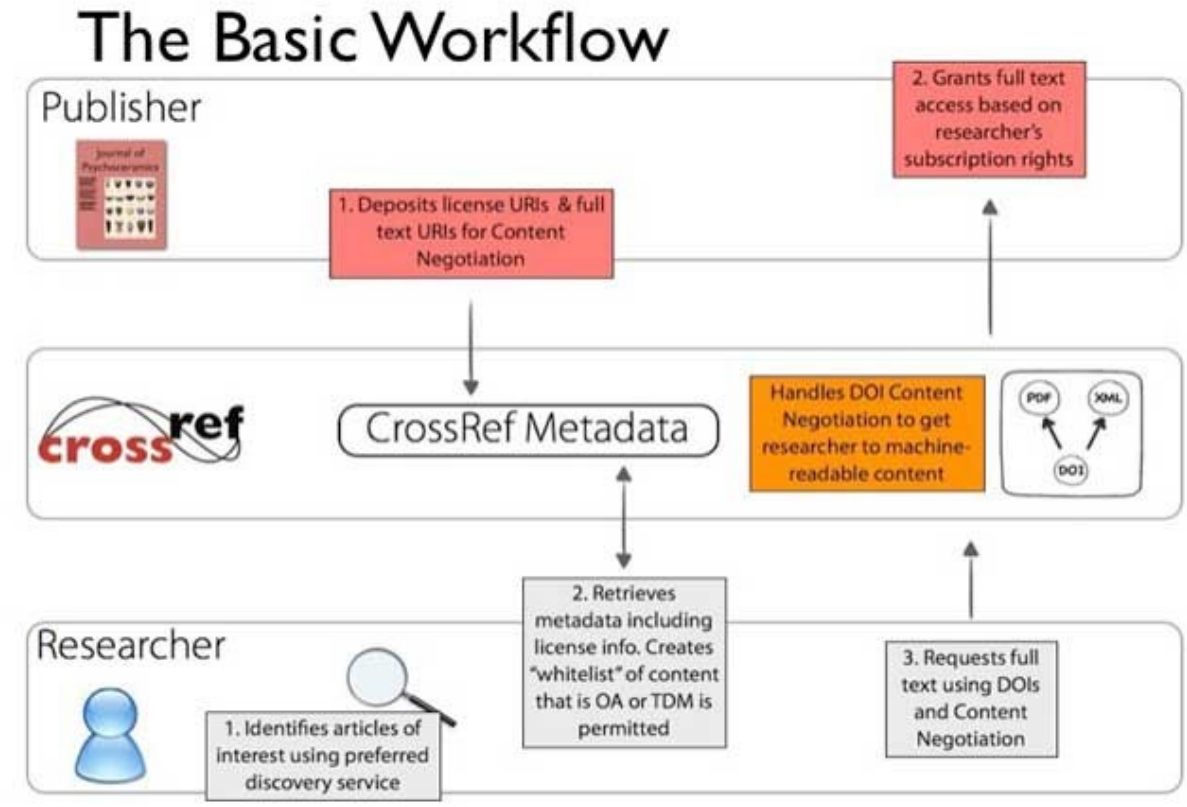


Figure 4.1: The basic CrossRef Text and Data Mining Services workflow [Lam15]

For this research, the articles and their DOIs are retrieved using previous methods, as in Step of the researcher's task in Figure 4.1. These DOIs are used to retrieve metadata from the Crossref database, as in Step 2.

4.6 Finding geolocations and utilizing metadata

From the metadata retrieved using Crossref, the names of authors and their affiliations are used. For each document, related metadata for the first author is retrieved and checked for an affiliation. This might be a company, institute, or other geographical location. The name of this affiliation is processed with the help of Google's geolocation application in order to retrieve a latitude and a longitude. These coordinates are further used with the help of the Geopandas software in Python, whereby they are extrapolated to a country.

Chapter 5

Results

In this chapter the data mining model will be presented and discussed. Each step will be explained to the how and why. After the data mining model, the use case results are handled. The use case of mining FAIR papers with the help of the data mining model are interesting and will show some interesting results.

5.1 The data mining model

The data mining model created to mine papers in a structured order contains six steps. Finding which papers dataset to use, identify the URL's where the paper resides, retrieving DOI's from the found websites, use the DOI's to retrieve meta-data, pre-process the metadata, find corresponding geolocations from the authors and finally process the meta-data to get analytics.

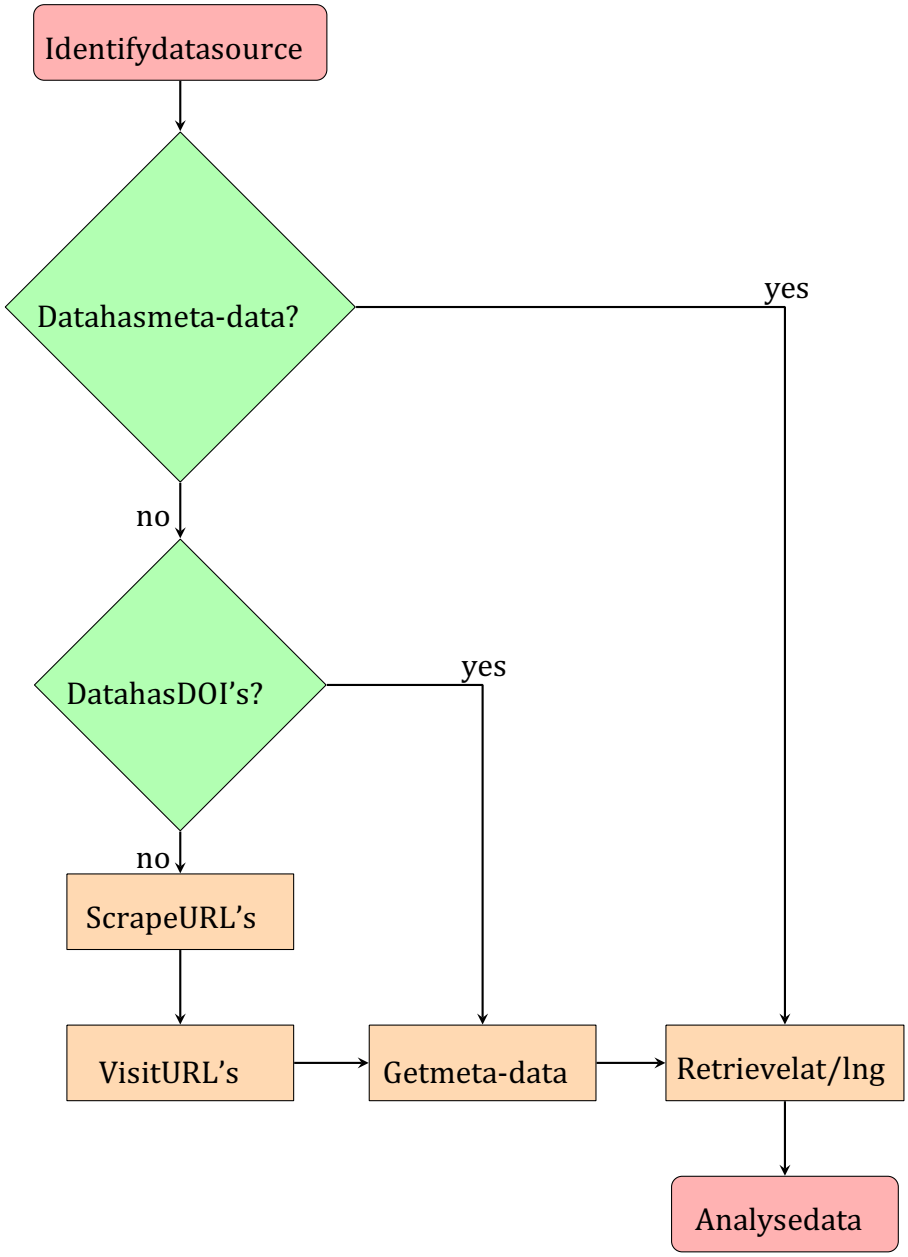
5.1.1 The dataset of papers

The first step of the model is to identify the dataset that will be used for mining paper related data. This can be any source as long as there are either webpages available where the paper is hosted or a list of DOI's. If the latter are available, one can skip the next 2 steps in the model. Such a dataset can be something like Google Scholar, Microsoft Acadamic or even specific publishers like Nature or Springer. As long as there are DOI's available or findable.

When the dataset is chosen the URL's for each individual paper needs to be identified and visitable. These URL's are essential for the next step where they are used to find DOI's. When your dataset already contains DOI's you do not have to identify the individual URL's.

5.1.2 Retrieving DOI's

This is an essential step in the model. The Digital Object Identifier's (DOI's) can link a paper to any kind of data. This DOI is intendent to be unique and only assigned to a single digital object. This identifier is utilised to retrieve more data from other sources, like CrossRef.



5.1.3 Obtaining meta-data

This step utilises the found DOI's from the previous step. We use it to find relevant (and non-relevant) data from a source that offers data about the DOI. One of those services is Crossref. Crossref stores

5.1.4 pre-processing meta-data

In this step the meta-data will be pre-processed in the sense that the unnecessary data is removed. Another approach is to only gather the meta-data that will be used in the next steps. Either way will result in only the specific meta-data that is required for the research.

5.1.5 Getting the geolocations

In order to find geolocations of papers, the filtered meta-data from the previous step is used. Depending on the approach the geolocations can be fetched. There is a choice to only look at the first author's affiliation. Another approach is to look at all affiliations and take the municipality, country or continent with the most affiliations.

5.1.6 Analysing the data

The final step is analysing the found data. Depending on the choice earlier made in the previous step, the data is analysed per municipality, country or continent. This last step fully depends on the research. For example, the funnelling of the geographical spread or an analysis year over year.

5.2 Results for the use case: FAIR paper usage

This section will describe the results gathered in the context of the use case. This use case aims to research the references to the original FAIR paper year over year and by country and continent. A side track of this use case is to look at publishers and how much meta-data they have shared.
The model of the first section in this chapter is used to gather and research the found data.

5.2.1 Finding data and DOI's

A total of 201 scholar pages were retrieved with up to 10 results per page. This is equivalent to a possible 2,010 papers. Eventually the titles and source urls of only 1,985 papers were collected. From these papers, some 1,320 DOIs were retrieved.

5.2.2 Meta-data retrieval and pre-processing

Through the service of Crossref, these DOIs were used to obtain the necessary metadata. This metadata contains authors and their affiliations. From this metadata a total of 327 geolocations were found. The results per website and a collection category of 'other' are displayed in Table 5.1.

As shown in the table, there are 10 individual resources that do not have any location affiliated with the author in the Crossref database. There are four resources that have less than 10% of the retrieved papers that have an affiliation stored in Crossref.

Four of the specifically tailored retrieved sources have a 100% score of authors with an affiliation in the Crossref database. Two others have a high, but not 100%, score for the number of affiliations submitted to Crossref.

Source page	# papers	# DOI's	# locations	% locations of papers
springer.com	158	157	0	0%
sciencedirect.com	139	137	0	0%
academic.oup.com	122	118	101	72.66%
nature.com	111	110	0	0%
onlinelibrary.wiley.com	87	81	81	93.10%
ieeexplore.ieee.org	68	68	0	0%
ncbi.nlm.nih.gov	63	57	2	3.18%

frontiersin.org	58	58	0	0%
researchgate.net	47	10	3	6.38%
biorxiv.org	45	43	0	0%
mdpi.com	44	44	0	0%
journals.plos.org	41	41	0	0%
arxiv.org	31	0	0	0%
dl.acm.org	24	24	24	100%
search.proquest.com	23	23	2	8.70%
tandfonline.com	22	22	22	100%
pubs.acs.org	19	19	19	100%
osf.io	18	0	0	0%
journals.sagepub.com	15	15	15	100%
mit.edu	5	5	2	40%
Other	845	278	56	6.63%
Totals	1985	1320	327	16.47%

Table 5.1: Counts of retrieved papers, DOI's and first author locations from different websites

5.2.3 Analysing the data

On the left side of the table, plots are shown where the papers were published by country. Every time a geolocation of the affiliation of the author is assigned to the country, the longitude and latitude are present. The legend shows the colour range and the corresponding number of papers represented. The date range is from 2016 to the end of 2020. Each time the number of papers of a country is added to the number of papers from the year before. This shows that a large number of papers have been written in the US.

On the right-hand side, the number of papers is displayed per continent. Due to the large number of papers assigned to individual countries, and especially the US, most other countries are next to indistinguishable. The aggregation per continent shows that a large number of papers mentioning FAIR principles have been written in the continents of North America and Europe (where Russia is also considered a part of Europe). Next best in terms of exposure are Asia and Oceania.

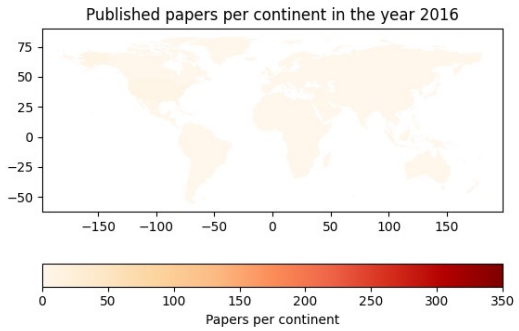


Figure 5.1: Papers from 2016 per country

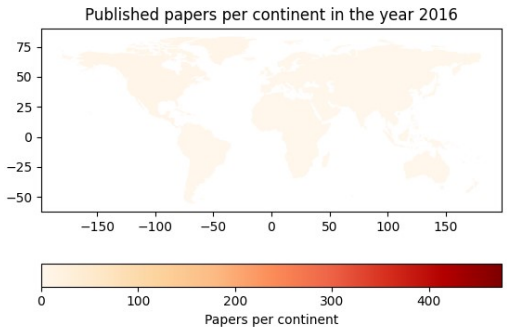


Figure 5.2: Papers from 2016 per continent

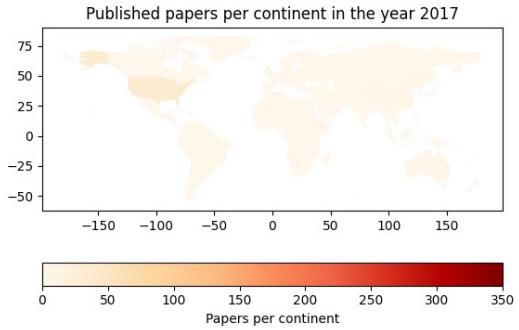


Figure 5.3: Papers from 2017 per country

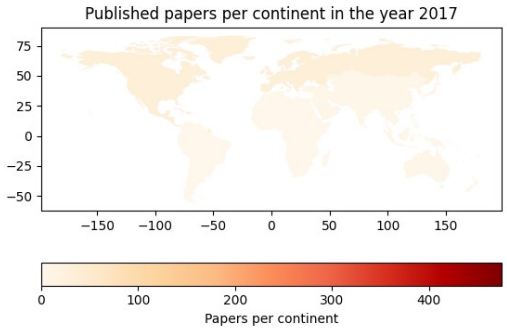


Figure 5.4: Papers from 2017 per continent

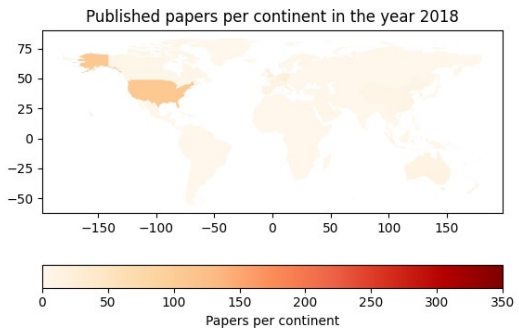


Figure 5.5: Papers from 2018 per country

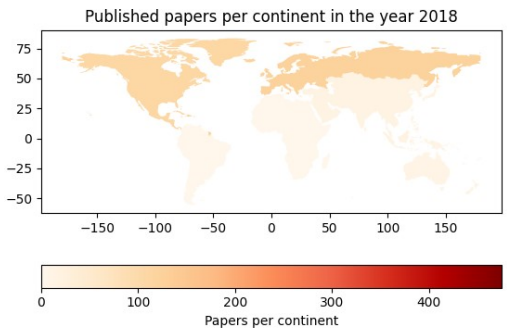


Figure 5.6: Papers from 2018 per continent

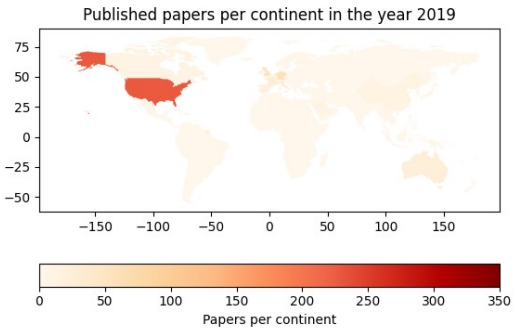


Figure 5.7: Papers from 2019 per country

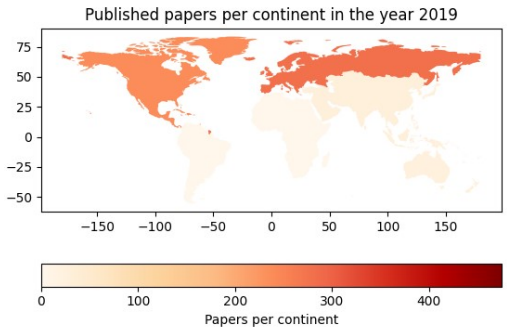


Figure 5.8: Papers from 2019 per continent

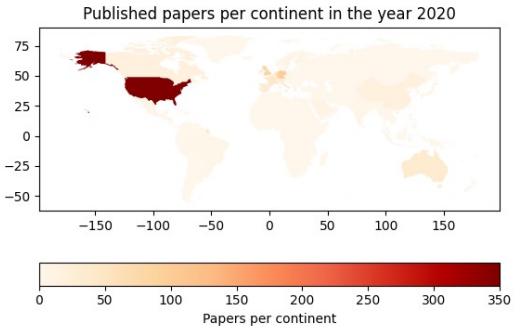


Figure 5.9: Papers from 2020 per country

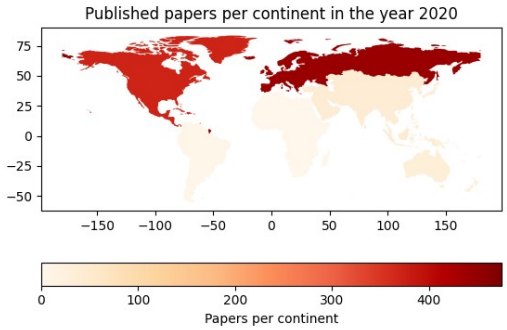


Figure 5.10: Papers from 2020 per continent

Chapter 6

Conclusion

The conclusion chapter consist of two parts. The first part will draw conclusions from the data mining model as described in the results (chapter 5.1). The second part will describe the conclusions drawn from the use case.

6.1 The full data mining model

The main goal of this research is to structure a data mining model that can be used to do large scale research to meta-data of published papers. That goal is researched with the help of a use case that requires such a model.
The final model consists of six steps, where two steps can be skipped if the dataset already contains DOI's and three steps can be skipped if the data source already contains meta-data. This shows that the dataset one chooses can be helpful in the process. But when you only have paper titles and URL's to the source of such a paper, there is a lot of work to be done. Like the use case in this research, where Google Scholar is used, a lot of manual steps need to be taken before the data is acquired where you want to draw your conclusions from. In the end this model will give handles to researchers to see how they can do a large scale research. It also shows that research, often created with public money, is not always as accessible and findable as one can expect.
The number of different services and tools required to get all necessary data is not small. It is not always clear what data is available and the quality of available data per paper varies as well.

6.2 The use case: FAIR paper usage

In this thesis the adoption of the FAIR framework in research publishing is investigated by mining papers that refer to the first FAIR paper. Its aim is to identify where FAIR principles are already being adopted and where the framework and its benefits could be better highlighted and promoted.

Results show which authors first wrote papers that refer to the original FAIR proposal by Wilkinson et al. Through data mining various resources, from Google Scholar to publishers' websites and Crossref, it is shown that most papers referring to FAIR are written in Europe and North America. The various plots from Chapter 5 show that initially the FAIR paper is most referred in the continent of North America, while after 2017 comparatively more papers referring to the FAIR principles are published in Europe.

Looking at the results from the Crossref database, there are major differences between publishers in terms of stored affiliations of paper authors. A couple of large publishers such as

Springer Publishing, *Science Direct* and *Nature* do not store such information. From all mined papers they score 0% of stored first author locations in the Crossref database. On the other hand, two large publishers, Oxford Academic and the Wiley Online Library, have a first author affiliation storage percentage of 72.66% and 93.10% respectively.

When analysing the results through the lens of planned behaviour theory we can conclude that in North America and Europe each year the amount of papers referencing FAIR is growing, suggesting that the attitude towards FAIR in these two regions is positively increasing. This also applies to the subjective norm, where the more that peers use the framework, the more researchers will use it as well. This attitude and subjective norm will result in the same behaviour: actually using it. The data shows that on the continents where it is used, the framework is referred to more and more.

This brings us to the sub-question, 'how easy is it to find data about FAIR and its adoption?' The answer is: not as easy as assumed. Google Scholar does not have an API that can be used by researchers to obtain metadata. A reasonable alternative is the meta-database of Crossref. However, this requires specific DOIs to find metadata and has no obvious search function, which Google Scholar offers. Combining the two is the next best solution.

Finally, we can answer the research question: where is FAIR being adopted and where is there need for more awareness of the FAIR framework in the world? As the data shows, Europe and North America have enough traction. FAIR usage is also steadily growing in Oceania, but not as much as in the aforementioned continents of Europe and Northern America.

6.3 Agenda setting

The model as described in chapter 5.1 will give a clear path to follow in doing research to large amounts of meta-data from papers where doing the same thing by hand will take humans to long. Combined with the use case as concluded in chapter 6.2 and the fact that the European Union requires the FAIR framework for researchers in order to get grants[5] displays the correctness of Kingdon's agenda setting model as described in chapter 3.3.

There is an obvious problem that a lot of research data is scattered across the globe. All this research is published, and just like the places where research is done and published, these papers are scattered around the internet. Finding relevant research for one's own research is cumbersome and labour intensive.
The FAIR framework provides a solid base and policy for researchers to share their data in a findable, accessible, interoperable and reusable manner. This framework is part of the policy stream as described by Kingdon.
A small step has been taken by the European Union - the politics stream - where FAIR is required in order to receive grants for researchers.

[5] https://www.dtls.nl/2016/04/20/european-commission-allocates-e2-billion-to-make-research-datafair/

All together this can lead to an agenda setting where politicians and civilians can use these principles to create better and durable places for researchers to share data and accelerate their research and discoveries.

Chapter 7

Further research and discussion

This research shows there is a lot of 'open' data hidden for the general public. Publishers do not always share their data and the quality of the data varies. One could add multiple data sources together or use multiple meta-data resources.

The research in this thesis shows that it is hard to make a model that fits all. Combining services could be an outcome to get more accurate results, but also results in more work to setup. The research also shows that not all publishers share meta-data from the scientific, mostly funded with public money, meta-data.

The use case shows that a lot of the resulting papers are published in Europe and the United states. This can raise other questions: Is this because of the publishers that were researched? Or is it due to the language selection? Is it maybe the use of only Google Scholar? Are more affiliations be taken into account when determining the location of the published paper? These are all questions that are worthwhile researching, but out of the scope of this research.

Bibliography

[KS84] John W Kingdon and Eric Stano. *Agendas, alternatives, and public policies*. Vol. 45. Little, Brown Boston, 1984.

[Qui86] Paul J Quirk. "" Agendas, Alternatives, and Public Policies", John W. Kingdon (Book Review)". In: *Journal of Policy Analysis and Management* 5.3 (1986), p. 607.

[CMS99] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. "Data preparation for mining world wide web browsing patterns". In: *Knowledge and information systems* 1.1 (1999), pp. 5–32.

[Pag+99] Lawrence Page et al. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[Cha00] Soumen Chakrabarti. "Data Mining for Hypertext: A Tutorial Survey". In: *SIGKDD Explor. Newsl.* 1.2 (Jan. 2000), pp. 1–11. issn: 1931-0145. doi: 10.1145/846183.846187. url: https://doi.org/10.1145/846183.846187.

[Jac05] P´eter Jacso´. "Google Scholar: the pros and the cons". In: *Online information review* (2005).

[Han07] David J Hand. "Principles of data mining". In: *Drug safety* 30.7 (2007), pp. 621–622.

[Pas10] Norman Paskin. "Digital object identifier (DOI®) system". In: *Encyclopedia of library and information sciences* 3 (2010), pp. 1586–1592.

[Kin11] John W Kingdon. "Agendas, alternatives, and public policies (Updated". In: *Glenview, IL: Pearson* 128 (2011), pp. 251–257.

[Lam15] Rachael Lammey. "CrossRef text and data mining services". In: *Insights* 28.2 (2015).

[Wil+16] Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (2016), pp. 1–9.

[Mon18] Barend Mons. *Data stewardship for open science: Implementing FAIR principles*. CRC Press, 2018.

[Wid18] Bastian Widyatama. "Applying Kingdon's Multiple Streams Framework in the Establishment of Law No. 13 of 2012 Concerning the Privilege of Yogyakarta Special Region". In: *Journal of Government and Civil Society* 2.1 (2018), pp. 1–18.

[Rei+20] Mirjam van Reisen et al. "Towards the tipping point for FAIR implementation". In: *Data Intelligence* 2.1-2 (2020), pp. 264–275.

[LLC21] MultiMedia LLC. *Scraper API Homepage subtitle*. 2021. url: https://www.scraperapi.com/ (visited on 07/26/2021).