



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Alternative Altmetrics

Jacob Hobbs

October 2019

Submitted as part of the assessment for ENGEN582-19Y
within
The School of Engineering

Declaration of Authorship

I, Jacob Hobbs, declare that this thesis, titled ‘Alternative Altmetrics’, and the work presented in it, are my own except as noted. I confirm that:

- This work was done wholly or mainly as part of ENGEN582 at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Acknowledgements

I would like to thank my supervisor, David Nichols.

Abstract

This document outlines the creation of a service for providing altmetrics. Altmetrics are an alternative to citation counts and use a count of mentions from non-academic sources. They can be advantageous due to their speed over citation counts. The existing altmetric providers have two main limitations, high cost and poor usability. An explanation of how identifiers for research items can be used to find mentions of those items, as well as an exploration of the web services available for collecting these identifiers and mentions is provided. Detailed is the creation of an ASP.NET web application that collects altmetrics. The application is both free to access and more usable than other free services. Altmetrics generated by the application are shown and the difficulties around a quantitative comparison with other services discussed. Results indicate that there may be a correlation between citation count and mention count, but this is not a focus. The limitations of this project centre around the cost of accessing certain APIs, especially Twitter, as well as the time required to implement the vast number of possible mention sources.

Contents

1	Introduction	6
1.1	Structure of Report	6
2	Background	8
2.1	Concepts	8
2.1.1	Identifiers	8
2.1.2	Identifier Formats	8
2.1.3	Mentions	9
2.2	Existing Providers of Altmetrics	9
2.2.1	Altmetric.com	9
2.2.2	Crossref Event Data	11
2.2.3	PlumX, Lagotto	11
2.3	Communication with Web Services	12
2.3.1	Application Programming Interfaces	12
2.3.2	JSON	12
2.3.3	Rate Limits	13
2.4	Summary	13
3	Design	14
3.1	Selected Identifier Sources	14
3.1.1	CORE	15
3.1.2	Unpaywall	15
3.1.3	DOI	16
3.1.4	shortDOI	17
3.1.5	Handle	17
3.1.6	Mendeley	18
3.1.7	PubMed	19
3.1.8	Scopus	20
3.2	Selected Mention Sources	20
3.2.1	Reddit	21
3.2.2	Twitter	21

3.2.3	Wikipedia	22
3.3	Excluded Mention Sources	22
3.4	Identifier Collector Prototype	23
3.5	Web Application	24
3.5.1	Endpoints	24
3.5.2	User Interface	25
4	Results	26
4.1	User Interface	26
4.2	Speed and Rate Limits	26
5	Discussion	31
5.1	Mention Counts	31
5.2	Limitations of Results	32
6	Conclusion	33
6.1	Future Work	33
	Bibliography	34
7	Appendices	36
7.1	Identifiers and Their Formats in Greater Detail	36
7.1.1	arXiv	36
7.1.2	CORE ID	37
7.1.3	DOI	37
7.1.4	shortDOI	37
7.1.5	Handle ID	37
7.1.6	Mendeley ID	38
7.1.7	PMID	38
7.1.8	PMCID	38
7.1.9	Scopus ID	38
7.2	Tables and Figures	39

Chapter 1

Introduction

Altmetrics seek to measure the attention received, or impact, of a scholarly work. The term altmetrics was coined by [1], who describe the disadvantages of traditional citation-based metrics as slow and with inaccuracies. The article proposes the concept of altmetrics as an alternative to citation-based metrics and details the tools available for its realization. Since the publication of this article, various altmetric tools have become available.

There are a number of existing services that provide a variety of altmetrics, but each has its issues. The aim of this project is to create an alternative to the currently available systems by developing a new system capable of providing altmetrics.

The effectiveness of altmetrics is still being researched, with results being contradictory and indecisive (see [2], [3] and [4]). The aim of this project, however, is to provide access to altmetrics, rather than attempt to prove their usefulness. Additionally, this project will focus on collection of online mentions as a form of altmetric, rather than offline sources or download counts.

1.1 Structure of Report

Chapter 2 provides a background to the key items involved in collecting altmetrics; identifiers for research items and online mentions. A number of existing altmetrics services are explored and important concepts involved in creating and interacting with web services defined.

Chapter 3 provides the evaluation process for the various web services available for collecting both identifiers and mentions. A description, as well as information key to implementing these services, is provided for the web services implemented in this project. Finally, the development of a prototype

for collecting identifiers, and of the final web application is summarized.

Chapter 4 provides images of the final application and its use, as well as information on its performance. Finally, chapter 5 explores the mentions collected by the application and the difficulties associated with evaluation.

Chapter 2

Background

2.1 Concepts

2.1.1 Identifiers

For the purpose of this project, an identifier refers to any piece of information that could be used to uniquely identify a work. Identifiers for collections of work, such as the ISSN, are irrelevant.

As an example, the main digital identifier for individual works is the Digital Object Identifier (DOI) [5]. The following string, `10.1038/nature.2014.14583`, is an example of a DOI. A DOI acts as a unique identifier for a work and is also associated with the current URL where the work is published, and acts as a permanent link. As the DOI is so prevalent, many online systems offer functionality around it, making the DOI a common identifier across most systems and an important one to collect.

As is typical with digital systems, each system that references a work will likely provide an additional, system specific identifier. As a result of this, each work typically has several identifiers. Discussion around a work may use any of that works available identifiers to reference it.

2.1.2 Identifier Formats

Each of the identifiers used to reference a work may additionally come in several different formats. An easily recognizable example is a download link to a PDF, for example `https://www.arcjournals.org/pdfs/ajng/v5-i1/3.pdf`. This link is unique enough to be used as an identifier for a work, but it does not account for all possible versions of it. A user could possibly use `https://`, `http://` or even omit the `http(s)://www.` section completely and still have a functioning URL.

10.1038/nature.2014.14583
doi:10.1038/nature.2014.14583
dx.doi.org/10.1038/nature.2014.14583
<http://dx.doi.org/10.1038/nature.2014.14583>
<https://dx.doi.org/10.1038/nature.2014.14583>

Figure 2.1: Different possible DOI formats

Continuing with the DOI as an example identifier; as it is used as a permanent link for the landing page of publications, it can itself be used as a URL. A few of the possible DOI formats as they appear around the web can be seen in figure 2.1.

For a full list of the identifiers used in this project, including a short description and possible formats, see section 7.1.

2.1.3 Mentions

The definition of a mention in an altmetric sense is as you might expect; whenever someone refers to a work, a mention is said to have occurred. These mentions can come from a variety of different sources, such as social media, blogs, patents or government documents. Figure 2.2 shows an example mention from Twitter, using a DOI to reference the work. For this project, the count of these mentions for any given work is the altmetric that a user would use to gain insight on its impact.

2.2 Existing Providers of Altmetrics

See table 7.1 for the mention sources supported by each provider.

2.2.1 Altmetric.com

Currently, the largest provider of altmetric tools is Altmetric (or Altmetric.com). Altmetric.com provides limited free tools, such as their altmetric badge [6]. The badge shows the count of a works mentions across sources (see figure 2.3) and is typically placed on landing pages for publications. Following the **See more details** link produces a page that provides additional information, such as mention demographics and lists each individual mention.

Access of additional tools provided by Altmetric.com, such as managing multiple papers or tools for institutions and publishers, requires payment [7].



ARC Journals @arcjournals · Jul 8

Comprehensive Evaluation of Neuromorphic Computing

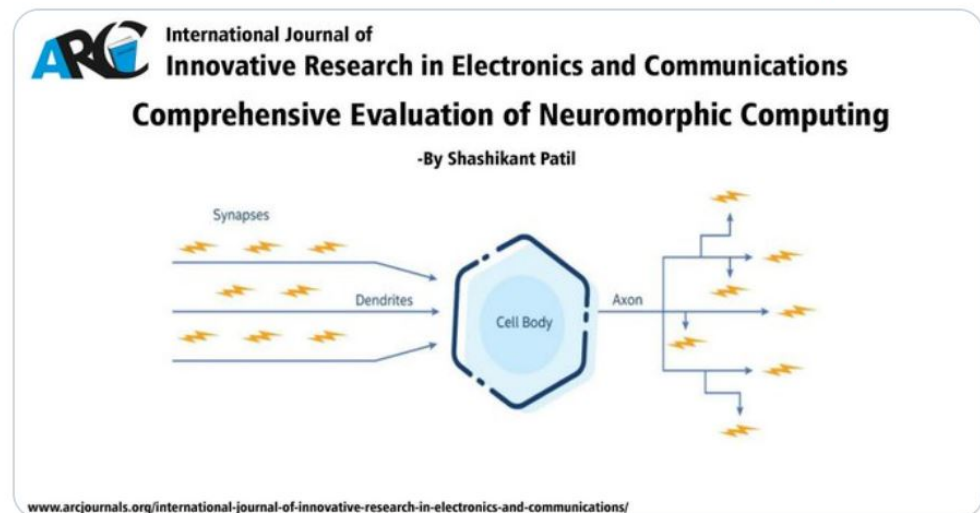
- By Shashikant Patil

DOI: [dx.doi.org/10.20431/2349-...](https://doi.org/10.20431/2349-...)

International Journal of Innovative Research in Electronics and Communications

arcjournals.org/international-...

#InnovativeResearch #Electronics #Communications #Neuromorphic



ARC Journals

Figure 2.2: An example of a mention from Twitter. <https://twitter.com/arcjournals/status/1148096133652025346>



Figure 2.3: An example Altmetric Badge for [8]. <https://www.altmetric.com/details/2068207>

The exact cost of these tools is not available on their website.

2.2.2 Crossref Event Data

Crossref Event Data is a free service operated by Crossref, a DOI registration agency. Event Data provides altmetrics in the form of raw data so that tools can be purpose built around it [9]. Data on its own is not very useful to a potential user (figure 7.1 gives an example of the provided data). Unfortunately, the Event Data service has issues which prevent it from being the ideal source of data for this project. As Crossref is a DOI registration agency, they do not support collecting mentions for works without a DOI [10].

2.2.3 PlumX, Lagotto

PlumX and Lagotto are alternatives to the above services. PlumX is an alternative to Altmetric.com and provides a free tool similar to the Altmetric Badge, as well as a full metrics suite as a product [11]. Lagotto is an alternative to Event Data. It is an open source project for providing altmetrics data but is in disrepair and has not been updated since 2016 [12]. It is mentioned as its documentation provides useful information around identifiers and mention sources [13].

2.3 Communication with Web Services

2.3.1 Application Programming Interfaces

Application programming interface (API) is a term used to describe a design pattern for allowing communication between applications [14]. This project focuses on the use of web services and as such, web APIs, to collect data. The web APIs involved in this project follow the representational state transfer (REST) architectural style. In essence, REST encourages the use of standard HTTP protocols to communicate between the client and the server.

An API call is said to have been made when a client requests the server to complete an operation. Each operation will usually have its own endpoint (or URL), as well as a set of optional parameters. Again, for the web APIs in this project, this call is made over HTTP.

As an example, the URL `http://api.eventdata.crossref.org/v1/events?obj-id=https://doi.org/10.1007/s00266-017-0820-4` makes a call to the Event Data API. The section `http://api.eventdata.crossref.org/v1/events` represents the events endpoint for this API. The section `obj-id=https://doi.org/10.1007/s00266-017-0820-4` represents a parameter and indicates which work we would like to receive events for using its DOI. As this call is made over HTTP, the result can be seen in a web browser by following the URL.

2.3.2 JSON

JavaScript object notation (JSON) is a format used to represent data [15]. All of the APIs used in this project offer JSON as a format for returning data. Figure 7.1 shows an example of the JSON returned by the Event Data service. In essence, JSON stores information in key-value pairs, with pairs structured in groups (or objects) and arrays.

In order to make use of the data contained in JSON, applications parse or serialize the JSON data into native objects. This functionality is usually provided through libraries for the various programming languages (for example ¹). Parsing and using JSON data requires knowledge of its structure. The documentation for an API will typically provide the JSON structure for each endpoint (for example ²). This example also shows how each API endpoint can return multiple different JSON structures depending on the set parameters and the different error cases.

¹<https://www.newtonsoft.com/json/help/html/Introduction.htm>

²<https://dev.elsevier.com/documentation/ScopusSearchAPI.wadl>

2.3.3 Rate Limits

Rate limits are often imposed by web services and their APIs. Rate limits set a cap on the number of calls specific users can make over a given period of time. This is done to prevent a single user from disrupting a system due to load. Rate limits are often provided in the documentation for an API (for example ³).

2.4 Summary

There are a variety of pre-existing tools for providing altmetrics, however they each have their disadvantages. The aim of this project is to create a free alternative to Altmetric.com and a more usable alternative to Crossref's Event Data. The system will make use of various web services to collect identifiers for a given work, format them, query mention sources and display the results as an altmetric to users.

³<https://developer.twitter.com/en/docs/basics/rate-limiting>

Chapter 3

Design

3.1 Selected Identifier Sources

Following is a description of each of the web APIs selected for collecting identifiers. Sources were evaluated based on the identifiers they provided and the ease at which they could be implemented. As the size of an online academic service decreases, the less likely it is to be used in a mention and the less use its internal identifier provides.

An example of an excluded identifier source is the Bielefeld Academic Search Engine (BASE).¹ Base is an aggregator and search engine for research. BASE requires an IP address to be registered for use with the API, which can be troublesome with the cloud hosting used for this project. Additionally, the documentation provided does not make the process of searching for identifiers apparent.² There are two other aggregator services, CORE and Unpaywall, that were chosen over BASE considering the time it would take to implement it.

Where information for the following APIs is given, it is sourced from that API's documentation, unless otherwise indicated. *Requires* indicates the identifiers supported by the API as an input, only one of these is required per call. *Provides* indicates the identifiers the API could potentially return to us, with those actually being returned dependant on what information the service contains. Section 7.1 provides additional information in individual identifiers and their formats.

¹<https://www.base-search.net/>

²<https://api.base-search.net/>

3.1.1 CORE

CORE is a research aggregator and collects open access research from a variety of different sources, such as repositories and journals (for a complete list, see ³). As a result of its aggregation from different services CORE aggregates the identifiers from these services. This makes it an ideal service for collecting identifiers. CORE provides no information for research that is not open access.

Documentation

<https://core.ac.uk/documentation/api/>

Endpoints

`core.ac.uk:443/api-v2/search/`

Requires: DOI, URL

Provides: CORE ID, DOI, PMID, Scopus ID

`core.ac.uk:443/api-v2/articles/get/`

Requires: CORE ID

Provides: DOI, URL

Authentication

CORE requires an API key appended as a parameter to every call made (for example `core.ac.uk:443/api-v2/search/?apiKey=12345`). An API key is available free for non-commercial purposes after registration.

Rate Limits

Free access entails a rate limit of five requests every 10 seconds. Rate limits for commercial applications are unknown and require contact with CORE.

3.1.2 Unpaywall

Unpaywall is an open access research aggregator. Access to their API is free and provides both the page URL and download URL for open access research from multiple sites hosting it. This is beneficial as services such as the DOI API will only provide the landing page for a research item, which will itself

³<https://core.ac.uk/data/providers/>

contain the download URL. With Unpaywall, download links do not have to be manually found from landing pages.

Documentation

<https://unpaywall.org/products/api>

Endpoints

api.unpaywall.org/v2/

Requires: DOI

Provides: URL

Authentication

No authentication required.

Rate Limits

There is a limit of 100,000 calls per day.

3.1.3 DOI

A DOI can be used as a URL when appended to dx.doi.org/, where the user will be redirected to that DOI's associated URL. The International DOI Foundation additionally provides an API that returns the URL associated with a DOI.

Documentation

<https://www.doi.org/factsheets/DOIProxy.html#rest-api>

Endpoints

doi.org/api/handles/

Requires: DOI, shortDOI

Provides: DOI, URL

When provided with a DOI, this endpoint returns the associated URL. When provided with a shortDOI, this endpoint will return the corresponding full DOI. This can then be sent to the same endpoint to receive the URL.

Authentication

No authentication required.

Rate Limits

No provided information.

3.1.4 shortDOI

The shortDOI API will provide the shortDOI for any given DOI. In cases where the given DOI does not have an associated shortDOI, one is created and returned.

Documentation

<http://shortdoi.org/>

Endpoints

shortdoi.org/

Requires: DOI

Provides: shortDOI

Authentication

No authentication required.

Rate Limits

No information provided.

3.1.5 Handle

The previously mentioned DOI API makes use of the Handle system, and as such, the Handle API behaves in much the same way. The Handle API is capable of returning the associated URL for both Handles and DOIs, whereas the DOI API does not support Handles. Although rate limits are not mentioned in the documentation, to balance load, DOIs are run only through the DOI API and Handles through the Handle API.

Documentation

https://www.handle.net/proxy_servlet.html

Endpoints

<hdl.handle.net/api/handles/>

Requires: DOI, Handle ID

Provides: URL

Authentication

No authentication required.

Rate Limits

No provided information.

3.1.6 Mendeley

Mendeley provides a number of academic tools.⁴ Their reference manager allows users to add and manage references or select from a catalog of previous references. Whenever a user adds a reference to their account, Mendeley updates the reference in their overall catalog to form a crowd sourced, canonical reference. Access to this catalog is available via API for free and is a useful source of identifiers. The URL returned by the API is URL of the item's Mendeley page.

Documentation

<https://dev.mendeley.com/methods/>

Endpoints

<api.mendeley.com/catalog>

Requires: arXiv, DOI, PMID, Scopus ID

Provides: arXiv, DOI, PMID, Scopus ID, URL

⁴<https://www.mendeley.com/guides/web>

Authentication

Mendeley uses OAuth authentication.⁵ After registering a Mendeley account, one must register an application which provides the application ID and application secret required for the Client Credentials Authorization Flow.⁶ These credentials are sent to the endpoint `api.mendeley.com/oauth/token`. If successful, the endpoint returns an access token which must be provided in the Authorization HTTP header of any API calls made. Access tokens expire after a period of time, which is indicated when the token is received.

Rate Limits

No information is provided in the documentation. A Tweet from the Mendeley API Twitter account suggests that the rate limit is 500 per hour.⁷

3.1.7 PubMed

PubMed is a search engine for research on medical topics and is managed by the US National Library of Medicine National Institutes of Health (NCBI).⁸ NCBI provides an API for converting between the PubMed identifiers and DOIs. This is important as many services that provide a PMID do not also provide the PMCID.

Documentation

<https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/>

Endpoints

`ncbi.nlm.nih.gov/pmc/utils/idconv/v1.0`

Requires: DOI, PMID, PMCID

Provides: DOI, PMID, PMCID

Authentication

No authentication required.

⁵https://dev.mendeley.com/reference/topics/authorization_overview.html

⁶https://dev.mendeley.com/reference/topics/authorization_client_credentials.html

⁷<https://twitter.com/mendeleyapi/status/47250684478889984?lang=en>

⁸<https://www.ncbi.nlm.nih.gov/pubmed/>

Rate Limits

No provided information.

3.1.8 Scopus

Scopus is a curated abstract and citation database.⁹ Payment is required for full access to Scopus, but the APIs used to provide identifiers can be accessed for free.¹⁰

Documentation

<https://dev.elsevier.com>

Endpoints

api.elsevier.com/content/search/scopus

Requires: DOI, PMID

Provides: DOI, PMID, Scopus ID

api.elsevier.com/content/abstract/scopus_id/

Requires: Scopus ID

Provides: DOI, PMID

Authentication

Scopus requires an API key sent as a parameter with each call made, similar to CORE. This is provided after registering an account and creating an API key.

Rate Limits

There is a limit of nine requests per second and an overall limit of twenty thousand requests per week.

3.2 Selected Mention Sources

Following is a short description of the mention sources selected for use in this project. Like the identifier sources, mention sources were evaluated on

⁹<https://www.elsevier.com/solutions/scopus>

¹⁰<https://dev.elsevier.com/about.html>

the amount of mentions they were likely to provide, as well as the ease of implementation.

3.2.1 Reddit

Reddit is a social media platform that facilitates discussion around user posted content, and so may contain discussion around research. Reddit provides free access to its API which can easily be used to query the entirety of Reddit.

Documentation

<https://www.reddit.com/dev/api/>

Endpoints

`api.reddit.com/search`

Requires: Query string

Authentication

Reddit uses OAuth authentication, which follows a similar process to the Mendeley API. A registered account and application are required.

Rate Limits

There is a limit of sixty requests per minute.

3.2.2 Twitter

From observation, Twitter arguably provides the most altmetric mentions of all the possible sources. Tweets come from researchers or other users interested in research and also institutions and researchers promoting their work. Twitter provides APIs for querying; however, results are limited to the past 7 days for free services.

Documentation

<https://developer.twitter.com/en/docs>

Endpoints

`api.twitter.com/1.1/search/tweets.json`

Requires: Query string

Authentication

Twitter uses OAuth authentication. A registered account and application are required.

Rate Limits

There is a limit of 450 requests per 15 minutes.

3.2.3 Wikipedia

Wikipedia provides citations to research in the References and Bibliography section of articles. Articles can be queried using the MediaWiki API.

Documentation

`https://www.mediawiki.org/wiki/API:Main_page`

Endpoints

`en.wikipedia.org/w/api.php?action=query`

Requires: Query string

Authentication

No authentication required.

Rate Limits

No set rate limits.

3.3 Excluded Mention Sources

There were a large number of mention sources excluded from this project. The majority of which due to the difficulty associated with implementing

them. Examples of mention sources supported by other altmetric providers can be seen in table 7.1.

An important mention source missed in this project is the Crossref Event Data service (see section 2.2.2). Event data on its own is a source of altmetrics; when provided with a DOI, it will return mentions for that work from a variety of different mention sources (again, see 7.1 for a complete list). Event Data could not be used as a mention source for this project as it became entirely unusable from 4 July 2019 onwards due to issues with load [16].

3.4 Identifier Collector Prototype

The first step in collecting altmetrics for a research item is to find as many of its identifiers as possible. To this end, a Python script was developed that, when provided with one or more identifiers of any type for a single item, returned a list of all found identifiers and their formats. Python was chosen due to the ease at which it is possible to make API calls and parse their responses.

First, the script contains a class `Identifier`, which defines identifier types and their formatting methods. Of note here is the concept of identifier importance. Each identifier type is given a weight based on the number of APIs that support it. Little testing was done as to the values of these weights and essentially, the DOI is given a high weighting due to its prevalence and all other identifiers given similar low weights.

The script contains a class for each of the previously mentioned identifier APIs. These contain lists of the required and provided identifiers previously mentioned. Each class also contains a function, `fetch_ids(Identifier)` which makes an API call with the provided identifier, parses the JSON result and returns a list of identifiers. For an example, the `shortDOI` class can be seen in figure 7.2.

Finally, the `program` class contains a list of all current collected identifiers, `collectedIdentifiers` and all API classes `availableAPIs`. The process begins by first selecting an API to use from `availableAPIs`. If `collectedIdentifiers` contains an identifier supported by an API, then it is available for use and is carried on to the next step. These APIs then have each identifier from `collectedIdentifiers` removed from their list of provided identifiers. If their list is now empty, then they cannot provide the program with any identifiers it does not already have and are removed from the selection process. The final API is selected by summing the weights of the identifiers remaining in each API's provided identifiers, the API that

provides the highest total is chosen. This API will likely provide the highest information gain, and so reduce the number of API calls required.

The API's `fetch_ids(Identifier)` function is called, and the results of the API call are added to `collectedIdentifiers`. The API is then removed from `availableAPIs` as it is unlikely to provide any new information if called again. The program now contains any additional identifiers returned, and one less available API. This process is repeated until all of the available identifier types have been found, or there are no more available APIs to use. Finally, each identifier has its list of formats generated and printed as output. An example output can be seen in figure 7.3.

3.5 Web Application

The development of a web service seemed to be the natural direction for this project to take, considering it is positioned as an alternative to other online services. The Identifier collector prototype was moved to an ASP.NET core web application, using Microsoft's C# language and .NET Core framework. The application is hosted on Microsoft's Azure App Service¹¹ and can be found at <https://alternativealtmetrics.azurewebsites.net/>.

Continuing on from the identifier collector prototype, the output list of formatted identifiers is combined into a single query string, which is then sent to each of the mention sources. The results returned are serialized into a JSON object, an example of which can be seen in figure 7.4. Additionally, the Crossref Metadata API is used to find the title and publication date of the item.¹²

3.5.1 Endpoints

- `.net/?handler=identifiers&doi=DOI`

This is an endpoint used to display the raw data for a single item where DOI is the DOI of an item. This is used to test the output of the application and provide results in JSON format instead of through the user interface. Remember that any identifier type can be used as a starting identifier, but for the sake of simplicity, only DOI has an endpoint at this stage.

- `.net/author/?name=NAME&dois=DOIS`

¹¹<https://azure.microsoft.com/en-in/services/app-service/>

¹²<https://github.com/CrossRef/rest-api-doc>

This endpoint is used to generate results for multiple items. **NAME** is the name associated with the list which could, for example, be the name of a researcher. **DOIS** is a comma separated list of DOIs for which the results will be generated. Currently these results are simply saved to the server as there was not enough time to develop and host a database approach.

- `.net/?name=NAME`

This endpoint is used to display results of previously generated lists, where **NAME** is the name of a list. The results are loaded into the user interface.

3.5.2 User Interface

The user interface was developed using the Bootstrap 4 design framework.¹³ The page uses the SB Admin template available for modification under the MIT license.¹⁴

The designed user interface follows the scenario of a researcher being provided altmetrics for their various publications by their institution. The institution would first generate altmetrics for the researcher using the endpoint mentioned above, then provide the researcher a URL to his or her dataset. A breakdown of the user interface shown to the researcher follows in chapter 4.

¹³<https://getbootstrap.com/docs/4.3/getting-started/introduction/>

¹⁴<https://startbootstrap.com/templates/sb-admin/>

Chapter 4

Results

4.1 User Interface

Figure 4.2 will be used to provide a walkthrough of the user interface design. The title labelled 1. displays the name of the current data set. The table labelled 2. displays the list of research items in the dataset along with the total count of mentions found for each one. Selecting a paper loads its content into the rightmost pane. In this pane, the graph labelled 3. gives a count of how many mentions have occurred per month. The pie chart labelled 4. shows the locations from which the mentions were collected. The table labelled 5. shows each individual mention and provides a URL so that its content can be viewed.

4.2 Speed and Rate Limits

Running the one hundred most cited items published by Harvard University in 2019 as sourced from Scopus took the application 17 minutes and 32 seconds or an average 10.5 seconds per item. The API with the most restrictive rate limit is likely the Mendeley API, which enforces a rate limit of five hundred calls per hour. At an average of one call every 10.5 seconds, the application will create around 342 calls each hour. This means that the application can conform to the various rate limits but running more than one item in parallel will likely exceed Mendeley's limit.

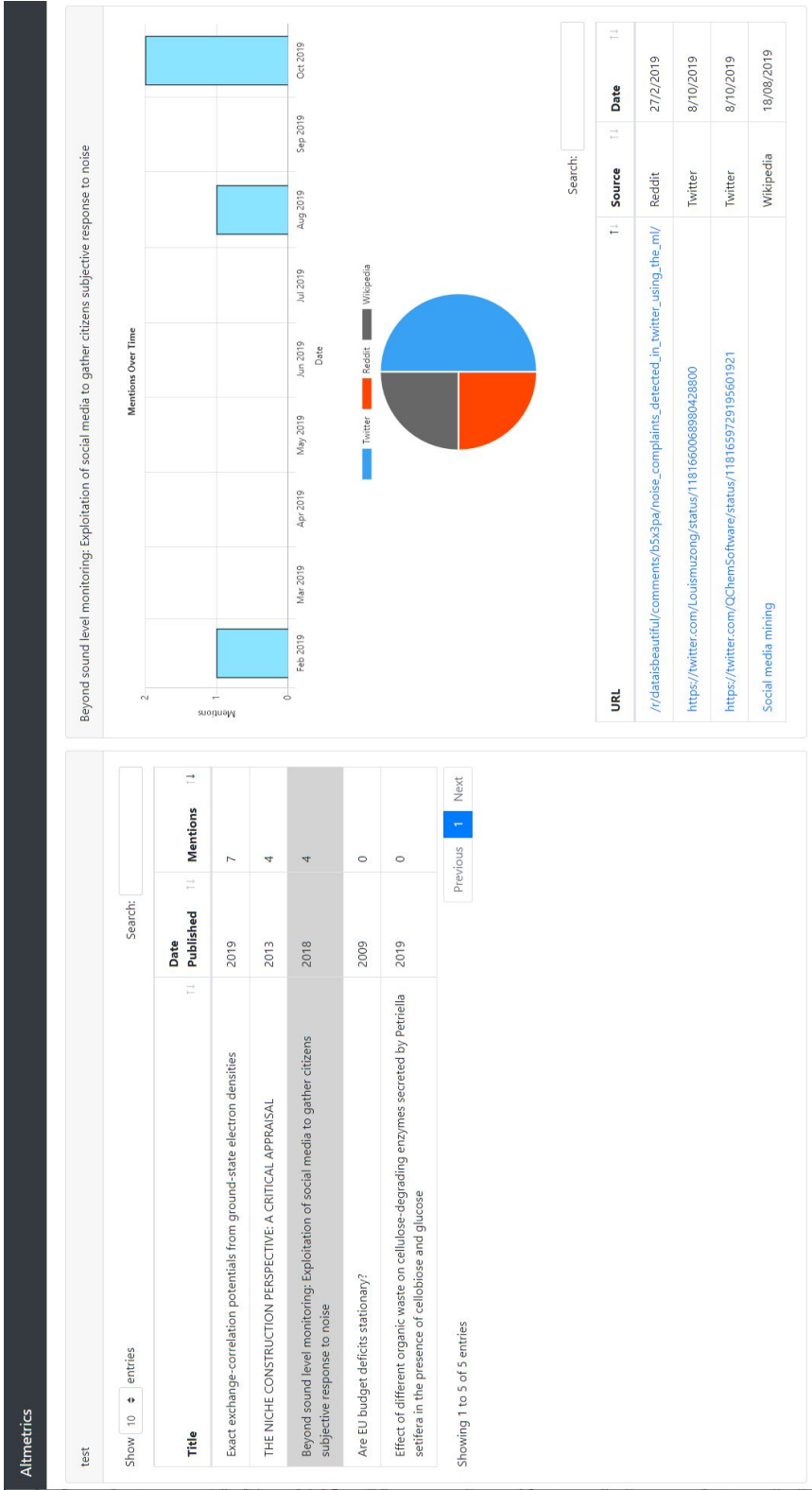


Figure 4.1: An example dataset loaded into the User Interface `https://alternativealtmetrics.azurewebsites.net/?name=test`

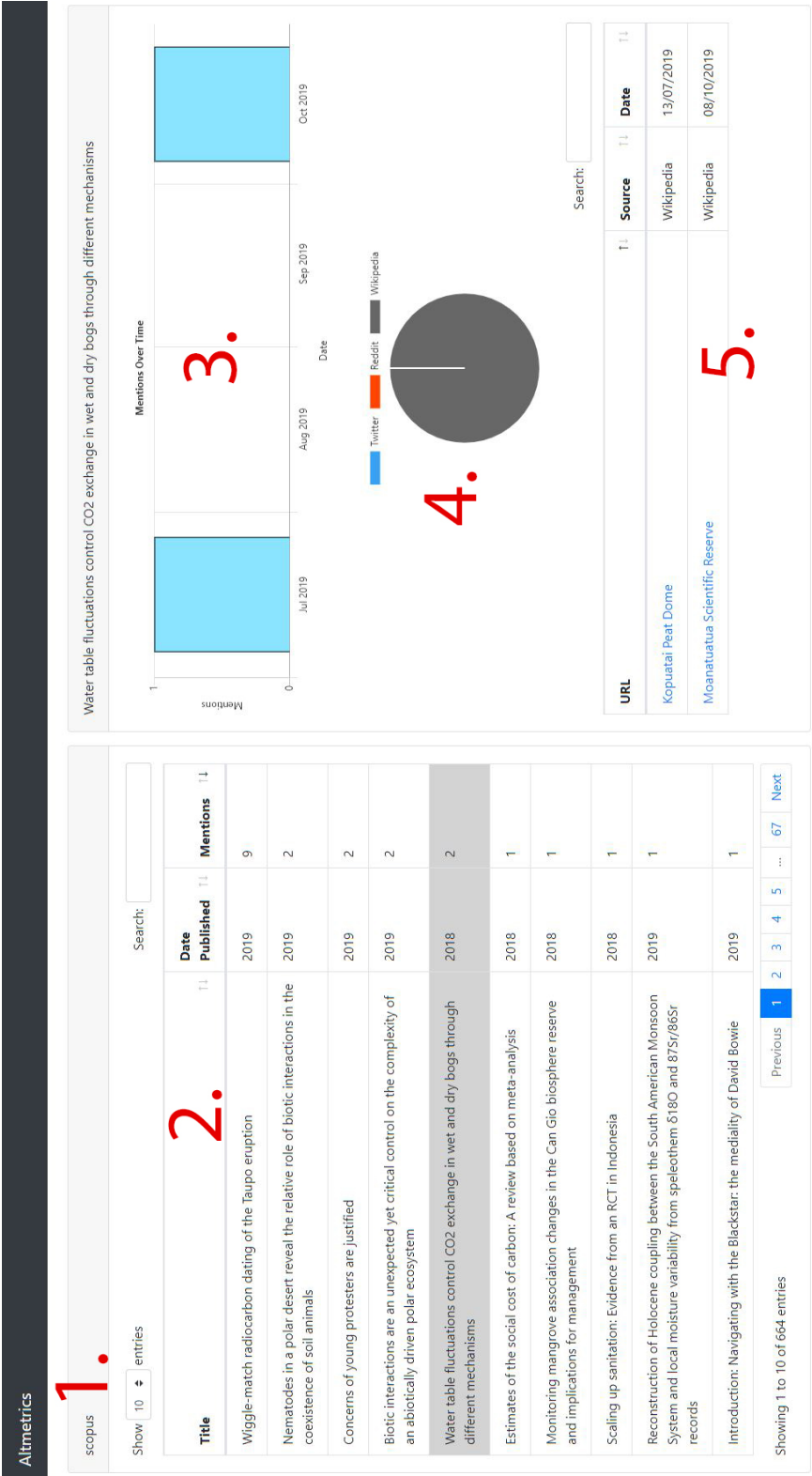


Figure 4.2: Annotated user interface, data sourced from 2019 University of Waikato publications sourced from Scopus <https://alternativemetrics.azurewebsites.net/?name=scopus>








Harvard		
Show	10 	entries
		Search: <input type="text"/>
Title  	Date Published  	Mentions  
Earth history and the passerine superradiation	2019	15
Large Fugitive Methane Emissions From Urban Centers Along the U.S. East Coast	2019	15
First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole	2019	11
First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole	2019	6
Impact of the WHO Framework Convention on Tobacco Control on global cigarette consumption: quasi-experimental evaluations using interrupted time series analysis and in-sample forecast event modelling	2019	6
Limbic-predominant age-related TDP-43 encephalopathy (LATE): consensus working group report	2019	4
Government policy interventions to reduce human antimicrobial use: A systematic review and evidence map	2019	4
The Pfam protein families database in 2019	2018	3
Highly structured homolog pairing reflects functional organization of the Drosophila genome	2019	3
First M87 Event Horizon Telescope Results. VI. The Shadow and Mass of the Central Black Hole	2019	2

Figure 4.3: 2019 publications from Harvard University with at least one citation as sourced from Scopus. Only the 10 highest mention counts are shown. <https://alternativealtmetrics.azurewebsites.net/?name=Harvard>

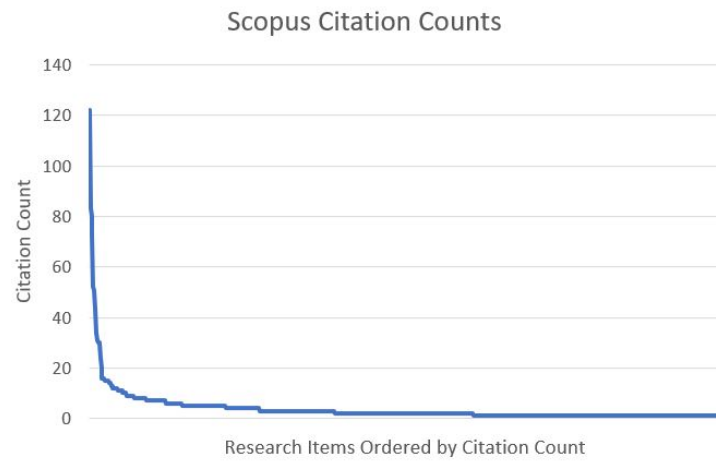


Figure 4.4: The citation counts of 2019 publications from Harvard University sourced from Scopus

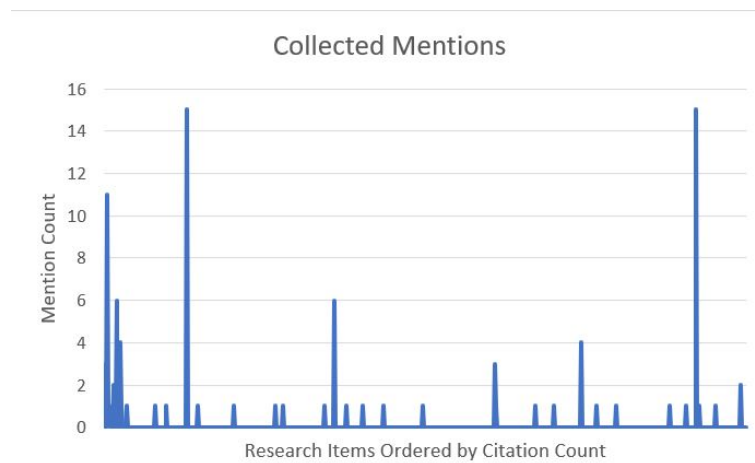


Figure 4.5: The count of collected mentions of 2019 publications from Harvard University as sourced from Scopus.

Chapter 5

Discussion

5.1 Mention Counts

Figure 4.3 shows the 10 most mentioned items published by Harvard University in 2019. Comparing with those from the University of Waikato in figure 4.2 provides the expected result of there being a greater number of mentions from the Harvard University publications. The most mentioned publication in the University of Waikato list is inflated, as it is recently published and thus had 9 Tweet mentions within the 7 days prior to this data being collected. This outlier is caused by the restrictions of the Twitter API, and removing it makes the greater popularity of Harvard Universities publications more apparent.

Figure 4.4 shows the citation counts as found on Scopus for the same 2019 Harvard publications. The x-axis represents each individual publication, however there are too many to label. The papers are ordered by citation count along the x-axis, with the left most items having a higher citation count, as seen by the plotted line. Figure 4.5 shows the count of collected mentions across all sources for the same data. Publications are distributed along the x-axis in the same order as figure 4.4. This graph provides a comparison between mention counts and citation counts.

There does appear to be a correlation between the two based on the group of publications with high mention counts on the left of figure 4.5, implying both a high mention and citation count. Again, the outliers with high mention counts further along the x-axis are likely recently published items with recent Twitter Mentions. Given the aim of this project is not to prove the validity of altmetrics, there will be no further investigation into the correlation between mention count and citation count, but these results are included as an interesting by-product of this project.

5.2 Limitations of Results

It is difficult to compare the mentions collected by this program with the other altmetric platforms quantitatively. The obvious candidate for comparison would be Crossref's Event Data, which provides easy free access to their mention data, but is unusable due to ongoing load issues (see section 3.3). Additionally, the free access tier to the Twitter search API means that this project only has access to the past 7 days' worth of Tweets. Given that the majority of mentions for research appear to come from Twitter, it becomes even more difficult to quantitatively compare output with Event Data. This is a resource limitation, as the Twitter API can become fully functional with little changes to this projects code if given the resources to pay for Twitters expensive enterprise APIs.

Chapter 6

Conclusion

The goal of this project was to create an alternative altmetrics service. The existing altmetrics services were discussed, and their shortcomings identified as cost and usability. Developed is a web application capable of providing altmetrics to users. It is free to access, and thus provides a cheaper alternative to Altmetric.com. By providing a user interface with graphical representations of data, it is a more usable alternative to Crossref's Event Data.

6.1 Future Work

The obvious limitation for this project when compared to other altmetric providers is the amount of mention sources. Realistically a piece of research could be mentioned almost anywhere, making the potential scope of this project almost limitless. The web services selected for use in this project had well defined APIs, simplifying their use. Companies like Altmetric.com and Crossref can utilize larger developer teams in order to integrate more difficult sources into their services, such as individually scraping blogs for mentions.

With more time also comes greater development of the user interface. Additional tools for multiple papers, researchers and institutions such as those provided by Altmetric.com. Refining a user interface is often a time intensive process, involving multiple iterations of user feedback and alteration that this project did not have the time for.

Bibliography

- [1] J. Priem and B. Hemminger, “Scientometrics 2.0: New metrics of scholarly impact on the social web,” *First Monday*, vol. 15, no. 7, 2010.
- [2] C. R. Sugimoto, S. Work, V. Larivière, and S. Haustein, “Scholarly use of social media and altmetrics: A review of the literature,” *Journal of the Association for Information Science and Technology*, vol. 68, no. 9, pp. 2037–2062, 2017.
- [3] S. Haustein, “Scholarly twitter metrics,” *CoRR*, vol. abs/1806.02201, 2018.
- [4] G. Eysenbach, “Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact,” *J Med Internet Res*, vol. 13, p. e123, Dec 2011.
- [5] International DOI Foundation, “DOI® Handbook.” <https://www.doi.org/hb.html>, Aug. 2016. (accessed: 23/10/2019).
- [6] Altmetric, “Altmetric Badges.” <https://www.altmetric.com/products/altmetric-badges/>. (accessed: 20/10/2019).
- [7] Altmetric, “Explorer for Institutions.” <https://www.altmetric.com/products/explorer-for-institutions/>. (accessed: 20/10/2019).
- [8] Merali, Zeeya, “Stephen Hawking: ‘There are no black holes’,” *Nature*, 2014.
- [9] Crossref, “Event Data.” <https://www.crossref.org/services/event-data/>, Sept. 2019. (accessed: 21/10/2019).
- [10] Crossref, “About the data.” <https://www.eventdata.crossref.org/guide/data/ids-and-urls/>. (accessed: 21/10/2019).
- [11] PlumX, “PlumX Metrics.” <https://plumanalytics.com/learn/about-metrics/>. (accessed: 22/10/2019).

- [12] Lagotto, “About Lagotto.” <https://www.lagotto.io/>. (accessed: 22/10/2019).
- [13] Lagotto, “Sources.” <https://web.archive.org/web/20181205222256/https://www.lagotto.io/docs/Sources>. (accessed: 24/10/2019).
- [14] Apigee, “Web API Design: The Missing Link.” <https://pages.apigee.com/rs/351-WXY-166/images/Web-design-the-missing-link-ebook-2016-11.pdf>.
- [15] json.org, “Introducing JSON.” <https://www.json.org/>. (accessed: 24/10/2019).
- [16] Crossref, “Continued Problems with load on the Event Data service..” <https://status.crossref.org/incidents/20rkc63xps5z>, July 2019. (accessed: 21/10/2019).

Chapter 7

Appendices

7.1 Identifiers and Their Formats in Greater Detail

This section gives a description of all the identifiers supported by this project, as well as each of their possible formats. Where formats are given, the first is an example of an unformatted identifier. Where these unformatted identifiers are not used in queries (too vague) they are italicised.

With the mention sources used in this project, the complete list of formats queried can be relaxed. All mention sources support the use of `url:` in queries to indicate a URL. Prepending a URL with `url:` allows the `http://` and `www.` section of URLs to be omitted, for example `url:doi.org/10.1038/nature.2014.14583`, reducing the required number of formats.

7.1.1 arXiv

arXiv is a repository for research that has not yet been published. The arXiv identifier represents a single version of an item.

Formats

1910.10752

`arxiv:1910.10752`

`arxiv.org/abs/1910.10752`

`arxiv.org/pdf/1910.10752`

7.1.2 CORE ID

The CORE ID is the internal identifier for individual items on CORE (see section 3.1.1).

Formats

17307675

`core.ac.uk/display/17307675`

`core.ac.uk/reader/17307675`

7.1.3 DOI

See section 2.1.1.

Formats

`10.1038/nature.2014.14583`

`doi:10.1038/nature.2014.14583`

`doi.org/10.1038/nature.2014.14583`

`dx.doi.org/10.1038/nature.2014.14583`

7.1.4 shortDOI

A shortDOI is a representation of a DOI in fewer characters. The two are often interchangeable, but some services will not support shortDOIs.

Formats

`10/aabbe`

`doi:10/aabbe`

`doi.org/10/aabbe`

7.1.5 Handle ID

A handle performs similarly to a DOI in that it is associated with a URL. Handle ID refers to the section after the `hdl.handle.net/` portion of a handle URL.

Formats

10289/10127

`hdl.handle.net/10289/10127`

7.1.6 Mendeley ID

A Mendeley ID is the internal identifier used to refer to a single item in the Mendeley catalog. This identifier is not used in any URLs, and so has no additional formats.

Formats

eaede082-7d8b-3f0c-be3a-fb7be685fbe6

7.1.7 PMID

PubMed ID (PMID), is an identifier provided by PubMed that refers to a single item indexed in PubMed.

Formats

`pmid:23193287`

`ncbi.nlm.nih.gov/pubmed/23193287`

7.1.8 PMCID

PubMed Central is a full text archive of research. The PubMed Central ID (PMCID) is used to refer to a single item.

Formats

`PMC3531190`

`ncbi.nlm.nih.gov/pmc/articles/PMC3531190/`

7.1.9 Scopus ID

A Scopus ID is the internal identifier used to represent a single item on Scopus. A Scopus ID cannot be used to create a URL for the page of a work, and so has no other formats.

Formats

80053651587

7.2 Tables and Figures

```
{
  "obj_id": "https://doi.org/10.1007/s00266-017-0820-4",
  "source_token": "45a1ef76-4f43-4cdc-9ba8-5a6ad01cc231",
  "occurred_at": "2017-02-27T19:06:32.000Z",
  "subj_id": "http://twitter.com/Ulcerasnet/statuses/83...",
  "id": "00001916-bbf6-4698-b8b8-26dbd885afa9",
  "action": "add",
  "subj": {
    "pid": "http://twitter.com/Ulcerasnet/statuses/83629...",
    "title": "Tweet 836291446168817665",
    "issued": "2017-02-27T19:06:32.000Z",
    "author": {
      "url": "http://www.twitter.com/Ulcerasnet"
    },
    "original-tweet-url": "http://twitter.com/UrgoTouch...",
    "original-tweet-author": "http://www.twitter.com/..."
  }
}
```

Figure 7.1: Example of a mention from Event Data, JSON format. <http://api.eventdata.crossref.org/v1/events?obj-id=https://doi.org/10.1007/s00266-017-0820-4>

Altmetric	Lagotto	PlumX	Crossref Event Data
Blogs	DataCite	Amazon	Blogs
F1000 Prime	PMC Citations	Blogs	Cambia Lens
Facebook	F1000 Prime	Github	DataCite
Media	Facebook	Goodreads	F1000 Prime
Mendeley	Journal Comments	News	Hypothes.is
Patents	Mendeley	Reddit	Media
Public Policy Documents	ORCID	Slideshare	Reddit
Publons	PMC Usage Stats	Sourceforge	Reddit Links
Pubpeer	Reddit	Stack Exchange	Stack Exchange Network
Reddit	Research Blogging	Vimeo	Twitter
Stack Overflow	ScienceSeeker	Wikipedia	Wikipedia
Twitter	Scopus	YouTube	WordPress
Web of Science	Twitter		
Wikipedia	Web of Science		
YouTube	Wikipedia		
	WordPress		

Table 7.1: The mention sources reportedly supported by existing altmetric providers. Altmetric.com: <https://help.altmetric.com/support/solutions/articles/6000136884-when-did-altmetric-start-tracking-attention-to-each-attention-source>
Lagotto: <https://web.archive.org/web/20181205222256/https://www.lagotto.io/docs/Sources>, PlumX <https://plumanalytics.com/learn/about-metrics/mention-metrics/>, Crossref: <https://www.eventdata.crossref.org/guide/sources/how-agents-work/>.

```

class ShortDoiApi():
    requires = [Identifier.DOI]
    provides = [Identifier.SHORT_DOI]

    @staticmethod
    def fetch_ids(id):
        #Make the API call
        response = requests.get("http://shortdoi.org/"
                                + id.value + "?format=json")
        #Parse the JSON into key-value pairs
        parsed = json.loads(response.content)
        #Find the returned identifier in the parsed JSON
        short_doi = parsed["ShortDOI"]
        #Return a new shortDOI identifier.
        return [Identifier(Identifier.SHORT_DOI, short_doi)]

```

Figure 7.2: Python code used to collect shortDOIs in the identifier collector prototype.

```

10.1038/nature.2014.14583
doi:10.1038/nature.2014.14583
url:doi.org/10.1038/nature.2014.14583
url:dx.doi.org/10.1038/nature.2014.14583
url:hdl.handle.net/11455/85791
url:larramendi.es/es/catalogo_imagenes/grupo.do?path=1021522
url:larramendi.es/es/catalogo_imagenes/grupo.do?path=1021521
url:larramendi.es/es/catalogo_imagenes/grupo.do?path=1021461
url:larramendi.es/es/consulta/registro.do?id=9757
url:mendeley.com/research/stephen-hawking-black-holes
url:nature.com/articles/nature.2014.14583
10/x25
doi:10/x25
url:doi.org/10/x25
url:dx.doi.org/10/x25

```

Figure 7.3: Output of the identifier collector prototype from starting identifier 10.1038/nature.2014.14583.

```

{
  "title": "Beyond sound level monitoring: Exploitation of social...",
  "date": "2018",
  "id": "10.1016/j.scitotenv.2018.12.071",
  "twitter": [
  ],
  "reddit": {
    "kind": "Listing",
    "data": {
      "after": null,
      "dist": 0,
      "facets": {
      },
      "modhash": "",
      "children": [
      ],
      "before": null
    }
  },
  "wikipedia": {
    "batchcomplete": "",
    "query": {
      "searchinfo": {
        "totalhits": 1
      },
      "search": [
        {
          "ns": 0,
          "title": "Social media mining",
          "pageid": 44518759,
          "size": 31691,
          "wordcount": 3230,
          "snippet": "Environment. 658: 69\u201319. Bibcode:2019ScTEn.658...",
          "timestamp": "2019-08-18T12:47:19Z"
        }
      ]
    }
  }
}

```

Figure 7.4: JSON output from web application <https://alternativealtmetrics.azurewebsites.net/?handler=identifiers&doi=10.1016/j.scitotenv.2018.12.071>