A methodology for gathering/annotating the raw-data/features of the documents citing a retracted article

Ivan Heibi[1], Silvio Peroni[1]

[1]University of Bologna

Ivan Heibi

ABSTRACT

Starting from a retracted article, we present a step-by-step methodology for gathering the raw-data of the documents which have cited such article (starting from its publication date) and define a strategy to annotate a set of features considering the collected data. The external services interrogated are all free and open. The methodology uses three external services: (a) OpenCitations COCI (http://opencitations.net/index/coci, used to retrieve citation data), (b) RetractionWatch database (http://retractiondatabase.org/, used to retrieve information of retracted articles), and (c) SCImago (https://www.scimagojr.com/, to retrieve subject areas and subject categories of publications). The methodology is divided into five steps: (1) identifying and retrieving the citing entities, (2) retrieving the citing entities characteristics, (3) classifying the citing entities according to subject areas and subject categories, (4) extracting textual values from the citing entities, and (5) annotating the in-text citations characteristics.
At the end of this process, we will have a dataset containing all the citing entities with all their data/features annotated in it. Starting from an empty dataset, each step of the methodology (from 1 to 5) enriches it with new variables.

EXTERNAL LINK

https://github.com/ivanhb/intext-cits-ret-method

EXTERNAL LINK

https://github.com/ivanhb/intext-cits-ret-method

PROTOCOL INFO

Ivan Heibi, Silvio Peroni . A methodology for gathering/annotating the raw-data/features of the documents citing a retracted article. **protocols.io**
https://protocols.io/view/a-methodology-for-gathering-annotating-the-raw-dat-bdc4i2yw

EXTERNAL LINK

https://github.com/ivanhb/intext-cits-ret-method

KEYWORDS

citation analysis, citation function, retraction, methodology

CREATED

Mar 07, 2020

LAST MODIFIED

Dec 07, 2020

PROTOCOL INTEGER ID

33916

GUIDELINES

The aim of this methodology is to build a dataset containing raw-data and information about the entities which have cited an examined retracted article. The methodology is divided into five steps which are summarized in the table below. For each step (row of the table) we mention its title (i.e. "Step" column), give a brief description (i.e. "Description" column), specify the inputs needed (i.e. "Input") and show the expected output (i.e. "Output" column). The output of each step extends the final dataset with new variables (columns) to it.

A	B	C	D
Step	**Description**	**Input**	**Output**
1) Identifying and retrieving the citing entities	Identifying the list of entities citing the retracted article and annotating their main metadata	DOI of the retracted article	*For each citing entity:* *1.1) DOI* *1.2) year of publication* *1.3) title* *1.4) venue id (ISSN/ISBN)* *1.5) venue title*
2) Retrieving the citing entities characteristics	Annotating whether the citing entities have been or have not been retracted as well	DOIs of the citing entities	*For each citing entity:* *2.1) is / is not retracted*
3) Classifying the citing entities according to subject areas and subject categories	Classifying the citing entities into areas of study and specific subject categories, following the SCImago classification	ISSN/ISBN of publication venues of citing entities	*For each citing entity:* *3.1) subject area* *3.2) subject category*
4) Extracting textual values from the citing entities	Extracting the citing entities' abstracts, the in-text reference pointers, citation contexts, title of the section where the in-text citations happen	DOIs of the citing entities	*For each citing entity:* *4.1) abstract* *4.2) in-text citation section* *4.3) in-text citation context* *4.4) in-text reference pointer*
5) Annotating the in-text citations characteristics	Annotating the intent and sentiment of each in-text citation, and specifying whether the text in citation contexts mentions the retraction of the cited article	In-text citations' contexts	*For each in-text citation:* *5.1) citation intent* *5.2) citation sentiment* *5.3) retraction is / is not mentioned*

An overview of all the steps needed for generating an annotated dataset for the citing entities of a retracted article. For each step, we give a brief description, the inputs needed, and the output. The output is represented as the expected list of annotated features that will enrich the final dataset.

All the related materials of the methodology are maintained in the dedicated Git repository at https://github.com/ivanhb/intext-cits-ret-method. The Git repository contains the directories:

- "*data/*": a results sample of each step of the methodology
- "*img/*": the images used in this document
- "*doc/*": other documents related to the methodology
- "*script/*": the Python scripts and Python Notebooks

Some of the Git repository contents have been also included and uploaded into this document. The ***method.py*** script is the main script for launching the methodology. On each step, we mention the correct command to execute using the main script. Along with the pure Python scripts (.py), we make also available a Python Notebook (.ipynb) defining the same operations: ***method.ipynb***.

ABSTRACT

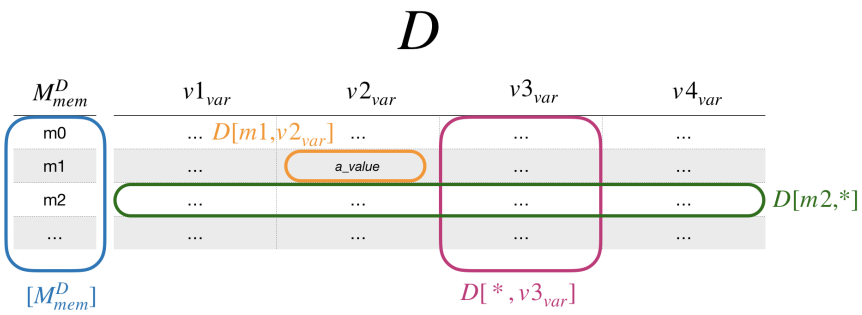
Starting from a retracted article, we present a step-by-step methodology for gathering the raw-data of the documents which have cited such article (starting from its publication date) and define a strategy to annotate a set of features considering the collected data. The external services interrogated are all free and open. The methodology uses three external services: (a) OpenCitations COCI (http://opencitations.net/index/coci, used to retrieve citation data), (b) RetractionWatch database (http://retractiondatabase.org/, used to retrieve information

of retracted articles), and (c) SCImago (https://www.scimagojr.com/, to retrieve subject areas and subject categories of publications). The methodology is divided into five steps: (1) identifying and retrieving the citing entities, (2) retrieving the citing entities characteristics, (3) classifying the citing entities according to subject areas and subject categories, (4) extracting textual values from the citing entities, and (5) annotating the in-text citations characteristics.
At the end of this process, we will have a dataset containing all the citing entities with all their data/features annotated in it. Starting from an empty dataset, each step of the methodology (from 1 to 5) enriches it with new variables.

BEFORE STARTING

This methodology takes for granted some basic knowledge regarding the scholarly publishing nature and the usage of the references and in-text citation styles.
Before starting, you need to make sure you have Python3.x installed on your computer, in addition, in order to correctly execute the Python-based scripts indicated on the methodology steps, you must install the required libraries defined in ***requirements.txt***. Please follow the official Python guidelines at https://wiki.python.org/moin/BeginnersGuide/ to check and eventually install python and the required libraries locally on your machine. Next, you need to download/clone the entire Git repository.
In the rest of this document we will use some common expressions summarize in the following glossary:

- "*value*": the **values** are written in italic surrounded by quotation marks. In case the **value** itself contains quotation marks, then the quotation marks are written in italic style too: *"value"*.
- $Dataset$: a **dataset** represented in a tabular format. The first letter of the dataset name is in uppercase.
- M^D_{mem} : **M** is the ***member*** of the **dataset *D*** (a dataset can have only one member); The first letter of the member's name is written in uppercase.
- $[M^D_{mem}]$: a set containing the **dataset's *(D) member (M)*** values.
- v_{var} : a ***variable* v.** The variable name is written in lowercase.
- $D[m, v_{var}]$: the ***value*** of the ***variable v*** for the corresponding ***member m*** of the dataset ***D*** (i.e. a table cell).
- $D[m, *]$: a set containing the ***values*** for a corresponding ***member m*** of the dataset ***D*** (i.e. a table row).
- $D[*, v_{var}]$: a set containing the ***values*** for a corresponding ***variable v*** of the dataset ***D*** (i.e. a table column).



Identifying and retrieving the citing entities

1 Starting from one retracted article identified with a DOI this step gets the metadata of all the citing entities included in the COCI dataset (the OpenCitations Index of Crossref open DOI-to-DOI references). We are only interested in a subset of attributes for the citing entities gathered. More specifically for each citing entity we want to annotate:

- The DOI value
- The year of publication
- The title of the article
- The ID of the venue (ISSN/ISBN)
- The title of the venue

In practical terms, this step will initialize our main dataset and include the above attributes in it. The next steps of this methodology will further enrich the same dataset with new variables that characterize each citing entity of the dataset.

Input: DOI of the retracted article

Output: creates the dataset $Cits\_Dataset$ with the initial variables/columns: doi_{var}, $title_{var}$, $year_{var}$, $source\_id_{var}$, and $source\_title_{var}$.

doi	title	year	source_id	source_title
...	...	...	...	...

1.1 First, we need to set the retracted article we want to examine. We consider articles that have officially received one or more retraction notice and had been eventually fully retracted. The Retraction Watch service reports and collects information about the retractions of scientific papers which they make available in an open queryable database at: http://retractiondatabase.org/. We use the Retraction Watch database to get the article we are interested in. Each record of the Retraction Watch database contains the following variables (columns):

1. Title, Subject(s), Journal, Publisher, Affiliation(s), Retraction Watch Post URL(s)
2. Retraction reasons
3. Authors
4. The Original Paper date/PubMedID/DOI
5. The retraction notice date/PubMedID/DOI
6. Article type(s) and the nature of the notice
7. Countries, If it is Paywalled? and Other notes

For the proceeding of this methodology, we consider the following variables from the above list:

1. The original DOI of the paper and its year of publication: although Retraction Watch reports the complete publication date we will only consider the year value
2. The retraction notice/s year: some articles might have more than one retraction notice, we will consider all these notices.

Another aspect we need to take into consideration at this stage is the in-text citation style. We need to take note and keep in mind this information, which will become very important in the next steps.

1. The DOI of the retracted article
2. The year/s of the retraction

Example:

1. 10.1016/S0140-6736(97)11096-0
2. 2004, 2010

1.2 Now we need to get the list of the entities citing the retracted article. We will query the COCI dataset (https://opencitations.net/index/coci). This dataset contains details of all the citations that are

specified by the open references to DOI-identified works present in Crossref (https://www.crossref.org/). OpenCitations provides a free APIs service to query and retrieve the COCI data at http://opencitations.net/index/coci/api/v1.
First, we get all the entities citing our seed article using the “citations” operation: http://opencitations.net/index/coci/api/v1#/citations/. Once we have the list of all the entities citing our retracted article, we will outline each citing entity with the following attributes: (a) the DOI value, (b) the year of publication, (c) the title of the article: (d) the ID of the venue (ISSN/ISBN), and (e) the title of the venue. These variables are part of the COCI datasets' metadata. To retrieve these values using the COCI APIs, we apply the “metadata” operation: http://opencitations.net/index/coci/api/v1#/metadata/, which requests the DOI/s value of the entity we are looking for and returns the metadata of such entity (if any). The COCI API does not necessarily have the metadata of all our DOI values (citing entities), in this case, such citing entities are excluded from our analysis.
In this step, we will initialize the dataset which will contain all our gathered/annotated data. The dataset will be extended with additional information/variables in the next steps. For the rest of this document, we will name our dataset $Cits\_Dataset$. This step's operations are done automatically by calling the following script.

Script to execute:
- *python3 method.py -s 1.2 -in <DOI>*

You can also specify a different output directory for the dataset to generate:
- *python3 method.py -s 1.2 -in <DOI> -out <DIR-PATH>*

Example:
- python3 method.py -s 1.2 -in "10.1186/1756-8722-5-31"
- python3 method.py -s 1.2 -in "10.1186/1756-8722-5-31" -out path/to/dir

A dataset containing all the entities citing the retracted article

Example:

doi	title	year	source_id	source_title
10.1002/jppr1998286443	Controversies in Vaccination with Measles, Mumps, Rubella (MMR) Vaccine	1998	issn:2055-2335; issn: 1445-937X	Journal of Pharmacy Practice and Research
10.1177/1362361398024006	Clinical Features Associated with Autism: Observations of Symptoms Outside the Diagnostic Boundaries of Autistic Spectrum Disorders	1998	issn:1362-3613; issn:1461-7005	Autism
...	...	...	...	...

step_1_2.csv

Retrieving the citing entities characteristics

2 To give the citing entities other attributes that aren’t part of the COCI metadata we need will use other services. The only thing we would like to check is whether any of the citing entities we are considering had been retracted as well. This value will be assigned to each citing entity of the $Cits\_Dataset$, so at the end of this step, we will have an extended version of the $Cits\_Dataset$ which embeds the additional $is\_retracted_{var}$. The first substep prepares the $Cits\_Dataset$, while the second substep shows how to correctly annotate the new variable.

Input: $Cits\_Dataset[*, doi_{var}]$

Output: extends the $Cits\_Dataset$ with the new variable $is\_retracted_{var}$

...	is_retracted
...	...

2.1 First, we need to prepare the $Cits\_Dataset$ for the upcoming annotation (done on the next substep). The dataset will be extended with the new variable, and its value set to "*todo*": $Cits\_Dataset[*, is\_retracted_{var}]$ = "*todo*". This operation is done automatically by calling the following script.

```
Script to execute:
python3 method.py -s 2.1 -in "<Cits_Dataset-Path>"

Example:
python3 method.py -s 2.1 -in "output/cits_dataset.csv"
```

An extended version of the $Cits\_Dataset$

Example:

doi	title	year	source_id	source_title	is_retracted
10.1002/jppr1998286443	Controversies in Vaccination with Measles, Mumps, Rubella (MMR) Vaccine	1998	issn:2055-2335; issn: 1445-937X	Journal of Pharmacy Practice and Research	TODO
10.1177/1362361398024006	Clinical Features Associated with Autism: Observations of Symptoms Outside the Diagnostic Boundaries of Autistic Spectrum Disorders	1998	issn:1362-3613; issn:1461-7005	Autism	TODO
...	...	...	...	...	...

step_2_1.png

2.2 To fill the new $is\_retracted_{var}$ we need to iterate over all the citing entities and manually verify whether any of the citing entities has been retracted as well. Again we use the RetractionWatch database (http://retractiondatabase.org/) and check all the citing entities using their DOI values. The "*todo*" values under the $is\_retracted_{var}$ are substituted with a yes/no value depending on whether the examined DOI does respectively has/hasn't received a final retraction.

The $CitsDataset$ with the annotated $is\_retracted_{var}$

doi	title	year	source_id	source_title	is_retracted
10.1002/jppr1998286443	Controversies in Vaccination with Measles, Mumps, Rubella (MMR) Vaccine	1998	issn:2055-2335; issn: 1445-937X	Journal of Pharmacy Practice and Research	no
10.1177/1362361398024006	Clinical Features Associated with Autism: Observations of Symptoms Outside the Diagnostic Boundaries of Autistic Spectrum Disorders	1998	issn:1362-3613; issn:1461-7005	Autism	no
...	...	...	...	...	...

step_2-2.csv

Classifying the citing entities according to subject areas and subject categories

3 The aim of this step is to annotate a subject area/s and a subject category/s for each citing entity in $Cits\_Dataset$. To do this we consider the venue identifiers (ISSN/ISBN) and classify them into specific subject area/s and a subject category/s using the SCImago Journal Classification (https://www.scimagojr.com/). This classification groups the journals into a subject area (27 major thematics), and subject category (313 specific subject categories). These values define two different levels: (1) a macro layer for the subject area, and (2) a lower layer for a specific subject category.
This step is divided into two main analyses. First, we focus on the citing entities having ISSN IDs, and then we move to analyze those having ISBN IDs. At the end of this step, the $Cits\_Dataset$ will be further extended with the two variables $area_{var}$ and $category_{var}$.
The first substep is a preparation phase. On substep 2 we handle the ISSN venues, while substep 3 and 4 analyze the ISBN venues. The final substep merges the results and populates the $Cits\_Dataset$.

Input: $Cits\_Dataset[*, source\_id_{var}]$

Output: extends the $CitsDataset$ with the new variables: $area_{var}$ and $category_{var}$

...	area	category
...	...	...

3.1 We first separate the ISSN and ISBN values into two datasets: $ISSN\_Dataset$ and $ISBN\_Dataset$. These datasets represent two indexes that include all the unique ISSN and ISBN values in the $Cits\_Dataset$. These two indexes are generated automatically using the script below. Both the datasets will have the $area_{var}$ and $category_{var}$ variables. The $ISBN\_Dataset$ contains the additional lcc_{var} variable (the reason will become clear

on substep 3.3). The following substeps will fill the corresponding values.

```
Script to execute:
python3 method.py -s 3.1 -in "<Cits_Dataset-Path>"

Example:
python3 method.py -s 3.1 -in "example_data/cits_dataset.csv"
```

ISSN_Dataset : a dataset containing the unique ISSN IDs of the *Cits_Dataset*

source_id	source_title	area	category
issn:0007-1048; issn:1365-2141	British Journal Of Haematology	TODO	TODO
issn:1356-1294	Journal Of Evaluation In Clinical Practice	TODO	TODO
...	...	...	...

step_3_1_1.csv

ISBN_Dataset : a dataset containing the unique ISSN IDs of the *Cits_Dataset*

source_id	source_title	lcc	area	category
isbn:9783319932231; isbn:9783319932248	Cognitive Errors And Diagnostic Mistakes	TODO	TODO	TODO
isbn:9783319932927; isbn:9783319932934	Autism In Translation	TODO	TODO	TODO
...	...		...	...

step_3_1_2.csv

3.2 This substep maps each unique ISSN value of our index *ISSN_Dataset* into its corresponding area and category following the SCImago journal classification. This process is done manually by checking each ISSN value using the SCImago Journal Rank service at https://www.scimagojr.com/. Among the returned information and metadata for the searched ISSN value, we write down the subject area and subject category. Journals might have more than one subject area or subject category, we will take into consideration all these values.
The following figure shows a result example from the Scimago Journal Rank service when searching for the ISSN value "*0273-9615*".

The subject area ($area_{var}$) and subject category ($category_{var}$) should be annotated inside the $ISSN\_Dataset$ following these rules:

- The ";;" segment (with white space at the end) is used as a separator between two different subject areas, and between two subject categories that belong to different areas.
- The ";" segment (with white space at the end) is used as a separator between two different subject categories that belong to the same area.

Consediring the above rules and the previous example (ISSN="*0273-9615*"), the correct form to annotate the $ISSN\_Dataset$ is:

- $area_{var}$: "*Medicine;; Psychology*"
- $category_{var}$: "*Pediatrics, Perinatology and Child Health;; Clinical Psychology; Developmental and Educational Psychology;;*"

An annotated $ISSN\_Dataset$ (the $area_{var}$ and $category_{var}$ are defined)

source_id	source_title	area	category
issn:0007-1048; issn:1365-2141	British Journal Of Haematology	Medicine	Hematology
issn:1356-1294	Journal Of Evaluation In Clinical Practice	Medicine	Health Policy; Public Health, Environmental and Occupational Health
...	...	...	...

step_3_2.csv

3.3 We need to classify also the ISBN venues into their corresponding subject areas and subject categories. Again we use the Scimago Journal classification. This choice is based on the fact that our aim is to have a standard pattern for all the venues regardless of their type (ISBN or ISSN).
The Scimago classification previously used for the ISSN sources belongs to the journal sources, therefore we can't apply a direct association of these values to the ISBN sources. We need a pre-elaboration which maps an ISBN classification model into the Scimago classification model (subject

area and subject category).
The ISBN classification model we used is the Library of Congress Classification (LCC, https://www.loc.gov/catdir/cpso/lcco/). First, we need to assign for each ISBN source in the $ISBN\_Dataset$ its corresponding LCC code. This operation is done manually using two main services: (a) the ISBNDB service (https://isbndb.com/), and (b) Classify (http://classify.oclc.org/classify2/), an experimental classification web service.
The LCC code values are written under lcc_{var} in the $ISBN\_Dataset$.

An updated version of the $ISBN\_Dataset$ having lcc_{var} annotated

source_id	source_title	lcc	area	category
isbn:9783319625416; isbn:9783319625430	Representing Scientific Knowledge	Q375	TODO	TODO
isbn:9783319638225; isbn:9783319638232	Recurrent Respiratory Papillomatosis	RC168.P15	TODO	TODO

step_3_3.csv

3.4 To compile the area and category of each ISBN source we call a function that maps the LCC codes to an area and category of the Scimago Journal classification. More precisely this function/algorithm will do the following operations for each member m of the $ISBN\_Dataset$:

1. Considers only the starting alphabetic segment of $ISBN\_Dataset[m, lcc_{var}]$ and find the corresponding LCC discipline using a pre-built lookup index. (e.g. "*RC360*" ->"*RC*" -> "*Medicine*")
2. Checks whether the value of the LCC subject is also a Scimago subject area using a pre-built Scimago index. If the corresponding value is present, the algorithm will automatically annotate the $ISBN\_Dataset[m, area_{var}]$ with such value, and the $ISBN\_Dataset[m, category_{var}]$ will have the same value with the addition of "*(miscellaneous)*" at the end of it, as it is done on the Scimago classification when denoting a Journal that treats general categories of a specific area. In case no corresponding Scimago area has been found the algorithm moves to point 3.
3. Checks whether the value of the LCC subject is a Scimago subject category using the same pre-built Scimago index. If the corresponding value is present, the program will automatically annotate the $ISBN\_Dataset[m, category_{var}]$ with such value, and the $ISBN\_Dataset[m, area_{var}]$ will have the same value used on the Scimago classification to denote the macro area of such category. In case no corresponding Scimago category has been found the program moves to point 4.
4. The program will annotate both $ISBN\_Dataset[m, area_{var}]$ and $ISBN\_Dataset[m, category_{var}]$ with the "*todo_manual*" value.

Once the above algorithm is done, we need to find the corresponding area and category for the records marked with the "*todo_manual*" value. To annotate such values we manually consult the complete LCC index (http://www.loc.gov/catdir/cpso/lcco/). The above algorithm is executed by running the following script:

```
Script to execute:
python3 method.py -s 3.4 -in "<CitsDataset-Path>"
```

```
Example:
python3 method.py -s 3.4 -in "example_data/cits_dataset.csv"
```

An annotated $ISBN\_Dataset$ (the $area_{var}$ and $category_{var}$ are defined)

source_id	source_title	lcc	area	category
isbn:9783319625416; isbn:9783319625430	Representing Scientific Knowledge	Q375	Computer Science	Computer Science (miscellaneous)
isbn:9783319638225; isbn:9783319638232	Recurrent Respiratory Papillomatosis	RC168.P15	Medicine	Internal medicine

step_3_4.csv

3.5 Finally, we merge the $ISSN\_Dataset$ and the $ISBN\_Dataset$ into our main $Cits\_Dataset$. The $Cits\_Dataset$ will be extended with the two variables $area_{var}$ and $category_{var}$. We might have multiple records in the $Cits\_Dataset$ having the same venue (ISSN/ISBN), the script will automatically assign the same value to all the duplicates. This process is done by calling the script below.

```
Script to execute:
python3 method.py -s 3.5 -in "<CitsDataset-Path>"

Example:
python3 method.py -s 3.5 -in "example_data/cits_dataset.csv"
```

The $Cits\_Dataset$ with the annotated $area_{var}$ and $category_{var}$

doi	title	year	source_id	source_title	is_retracted	area	category
10.1002/jppr1998286443	Controversies in Vaccination with Measles, Mumps, Rubella (MMR) Vaccine	1998	issn:2055-2335; issn: 1445-937X	Journal of Pharmacy Practice and Research	no	Medicine	Pharmacology (medical);;
10.1177/1362361398024006	Clinical Features Associated with Autism: Observations of Symptoms Outside the Diagnostic Boundaries of Autistic Spectrum Disorders	1998	issn:1362-3613; issn:1461-7005	Autism	no	Psychology	Developmental and Educational Psychology
...	...	...	...	...	...	...	...

step_3_5.csv

Extracting textual values from the citing entities

4 In this step, we enrich the $Cits\_Dataset$ with new variables that denote some textual values contained in the citing entities' full-text. The values we are interested in are:

1. The abstract ($abstract_{var}$): the abstract of the citing article in case there is any.
2. The in-text citation context/s ($intext\_citation.context_{var}$): the textual context/s that contains a reference-pointer (i.e. citation) to the retracted article
3. The in-text citation section/s ($intext\_citation.section_{var}$): the section/s that contains the reference-pointer (i.e. citation) to the retracted article
4. The in-text citation pointer/s ($intext\_citation.pointer_{var}$): the actual reference-pointer used (e.g. Heibi (2019))

The first substep prepares the $Cits\_Dataset$ to be filled later with the above values. Substep 4.2 discusses each one of the above values and describes how to correctly annotate them.

Input: $Cits\_Dataset[*, doi_{var}]$

Output: extends the $Cits\_Dataset$ with the new variables: $abstract_{var}$, $intext\_citation.section_{var}$, $intext\_citation.context_{var}$ and $intext\_citation.pointer_{var}$.

...	abstract	intext_citation .section	intext_citation .context	intext_citation .pointer
...	...	...		...

4.1 This substep extends the $Cits\_Dataset$ with the new new variables: $abstract_{var}$, $intext\_citation.section_{var}$, $intext\_citation.context_{var}$ and $intext\_citation.pointer_{var}$. The default value assigned to these fields is "*todo*".
The next substeps explain how to correctly replace and fill in the correct values. This substep process is made automatically by calling the following script:

```
Script to execute:
python3 method.py -s 4.1 -in "<Cits_Dataset-Path>"

Example:
python3 method.py -s 4.1 -in "example_data/cits_dataset.csv"
```

doi		abstract	intext_citation.section	intext_citation.context	intext_citation.pointer
10.1002/jppr1998286443	...	TODO	TODO	TODO	TODO
10.1177/1362361398024006	...	TODO	TODO	TODO	TODO
...		...	...	...	...

step_4_1.csv

4.2 To annotate the new variables we need to consult the citing entities' full-texts. Some of these are open and freely accessible, others are closed by paywalls. We consider only the entities that we had successfully found their full-text, all the others are removed from the $Cits\_Dataset$ and not considered. Finding the full-texts and removing the citing entities are two operations that should be made manually (for each citing entity in the $Cits\_Dataset$).

Once we have collected all the full-texts, we need to replace the "*todo*" values with the true corresponding values following the rules below:

The abstract ($abstract_{var}$):

Copy the entire abstract from each citing entity's full-text. In case no abstract has been found, we write an empty string. Examples of documents with no abstracts are book chapters or editorials.

The in-text citations pointer ($intext\_citation.pointer_{var}$):

To correctly annotate this variable we need to have a good background on the citing formats and how the reference pointers in the text are written. Look at the following guidelines: https://tinyurl.com/vtdd6x2 for a brief background on this topic.
We search inside the citing entities' full-text for all the in-text citations referring to our retracted article, and we write down the value used to point to the retracted article reference entry. For instance, this means that for a member m in $Cits\_Dataset$, the value of $Cits\_Dataset[m, intext\_citation.pointer_{var}]$ might be: *"Heibi(2019)"*.
Note that this value is the same one adopted for each in-text citation inside the document.

The in-text citations context ($intext\_citation.context_{var}$):

We want to write down the context of each detected in-text citation. We define our in-text citation context as the sentence that includes the in-text citation pointer (anchor sentence), plus the prior and the following sentence. There are some special cases we need to handle. If the in-text citation pointer:

1. Appears in a title: the context equals the entire title.
2. Appears in a table cell: the context equals the entire table cell.
3. Appears in the first sentence of a section/sub-section: the context equals the anchor sentence plus the sentence after.
4. Appears in the last sentence of a section/sub-section: the context equals the anchor sentence plus the prior sentence.

We might have more than one in-text citation in one citing entity, in this case, we must include the ";;" segment as a separator between every two different contexts. For instance, for a record m in $Cits\_Dataset$, the value of $Cits\_Dataset[m, intext\_citation.context_{var}]$ might be:

"*We will talk about this aspect later. As it was also observed in Heibi(2019). Now we move to the second point of our analysis. ;; This work takes into consideration new features. We are working on extending the previous work of Heibi(2019)*"

The in-text citations section ($intext\_citation.section_{var}$):

Each in-text citation is linked to the section where it occurred. If the related citing entity's full-text does not include any section/paragraph (e.g. an editorial), then the value of $Cits\_Dataset[m, intext\_citation.context_{var}]$ equals "*none*".

Otherwise, the in-text citation section is annotated using one/both these values:

1. **Type:** could be equal to one of the following values: (a) "*abstract*", (b) "*introduction*", (c) "*background*", (d) "*results*", (e) "*method*" and (f) "*conclusion*". We chose one of these values only if it is clearly inferred from the section title (e.g. the title contains the typology name). In case we can't link the section to any of these types we will take note of its position in the document: (a) "*first section*": appears in the first section of the article, (b) "*final section*": appears in the last section of the article, and (c) "*middle section*": it appears neither in the first section nor the final section.
2. **Title:** in case we have assigned to the section one of the previous 5 typologies, we will omit this value. Otherwise, we will annotate the exact title surrounded by brackets (*"..."*). So the value of $Cits\_Dataset[m, intext\_citation.context_{var}]$ will contain the title alongside the section position annotated on point(1).

We might have more than only one in-text section in each examined citing entity, in this case, we must include the ";;" segment as a separator between two different sections, and we use the "," as a separator between the type and title of the section.

For instance, this means that for a record m in $Cits\_Dataset$, the value of $Cits\_Dataset[m, intext\_citation.section_{var}]$ in case it contains two in-text citations is:
"*introduction;; final section; "Discussion" "*

doi		abstract	intext_citation.section	intext_citation.context	intext_citation.pointer
10.1002/jppr1998286443	...	Many people remember the extensive adverse publicity surrounding pertussis vaccine in the 1970s which led to public concern about the safety of vaccines. Subsequently, follow-up indicated no causal link between pertussis vaccination and brain damage although the risk of recurrent collapse (shock-like syndrome or so-called hypotonic-hyporesponsive episode) with a benign outcome, after such vaccination is higher than the background rate	none	In the United Kingdom from 1976-88 there were three major pertussis epidemics accounting for over 300 000 notifications and at least 70 deaths. 4 A recent 'early report' linked MMR vaccination with non-specific colitis and pervasive developmental disorder in children. 5 The suggestion that measles infection is causally linked to Crohn's disease had been disproved already6 but there has since been a decrease in MMR vaccination uptake with the potential for future morbidity. 7;; The suggestion that measles infection is causally linked to Crohn's disease had been disproved already6 but there has since been a decrease in MMR vaccination uptake with the potential for future morbidity. 7 The basis for the early report was 12 children whose parents had reported the onset of bowel disease and autism soon after MMR vaccination. 5The age when these conditions first appear however, is also the age for MMR vaccination. ;; The publication of the 'early report' based on only 12 patients resulted in a prodigious correspondence which canvassed the rift which can exist between clinical medicine and public health, 12 and the inadequacy of evidence related to problems of methodology. 11-21 The controversy in that correspondence was whether or not the paper byWakefield5 should have been accepted for publication. It attracted mass media attention and generated considerable public concern despite the lack of evidence linking MMR vaccination with the conditions suffered by the 12 children.	5
10.1177/1362361398024006	...	Autism is traditionally viewed as a pervasive developmental disorder, diagnosed on the basis of clinically observed behavioural symptoms in the absence of any quantifiable biochemical or neurological markers. Anecdotal reports of the coincidence of certain physiological anomalies in subgroups of children with autistic spectrum disorders, outside the syndrome's diagnostic definition, have not as yet been adequately catalogued. Findings are presented from parental interviews (n = 46), examining the type and prevalence of these additional health and psychological factors, and their connection to a possible immunological/metabolic basis to autism and associated spectrum disorders.	final section; "Discussion";; final section; "Discussion"	The factors that were revealed, however, indicate the need for further investigation into the possible links between the autistic syndrome and the prevalence of physiological factors. Speculation about a possible interplay between autism and various biochemical systems of the body, such as the immune system (Warren et al., 1996; Van Gent et al., 1997; Wakefield et al., 1998), suggests that specific behaviours associated with the autistic syndrome could also be accompanied by specific physiological problems associated with affected systems of the body. The findings that nearly two-thirds of Children showed a history of eczema or similar skin complaints, and over a quarter had problems with asthma—related conditions (both exceeding the general population incidence figures), imply the involvement of the immune system in some fashion.;; The incidence of migraines is compatible With the phenyls ulphotransferase enzyme deficiency hypothesis (Waring et al.,1996), which has also been associated with subsequent physiological difficulties such as problems with intestinal permeability (Waring and Ngong, 1993; Shattock and Savery, 1996). The significance of a link between parental and sibling health problems (especially related to problems of the immune system) and the presence of physiological symptoms coexisting with the overt diagnostic definition of autism supports the View of autism being an organic disorder, either immunological or metabolic (Shattock and Savery, 1996 ; Wakefield et al., 1998), with a genetic basis at least, in a subgroup of children.	Wakefield et al., 1998
...		...	...	...	...

step_4_2.csv

Annotating the in-text citations characteristics

5 In the previous step, we have enriched our $Cits\_Dataset$ with all the in-text citations we have in the citing entities' full-text. Yet, we still have not made any further content analysis on the context of the in-text citation. In this step, we add to the $Cits\_Dataset$ 3 characteristics correlated with the in-text citation/s:

1. $intext\_citation.intent_{var}$: the citation intent/reason/function is defined as the author's reason for citing a specific paper (e.g. because the citing document want to use the method defined in the cited paper),
2. $intext\_citation.sentiment_{var}$: the author's sentiment regarding the cited entity. We check whether the author's perception of the cited entity is positive/negative/neutral.

3. $intext\_citation.ret\_mention_{var}$: check whether at least one of the in-text citations contexts of a specific citing entity does explicitly mention the fact that the cited entity is retracted.

In the first substep, we will first prepare the $Cits\_Dataset$ to be filled with the above variables. Next, we move on describing how to correctly annotate them.

Input: $Cits\_Dataset[*, intext\_citation.context_{var}]$

Output: extends the $Cits\_Dataset$ with the new variables: $intext\_citation.intent_{var}$, $intext\_citation.sentiment_{var}$, and $intext\_citation.ret\_mention_{var}$.

...	intext_citation .intent	intext_citation .sentiment	intext_citation .ret_mention
...	...	...	...

5.1

This substep extends the $Cits\_Dataset$ with the new variables: $intext\_citation.intent_{var}$, $intext\_citation.sentiment_{var}$, and $intext\_citation.ret\_mention_{var}$. The default value of these variables is "*todo*".

The next substeps explain how to correctly replace and fill the correct values. This process is made automatically by calling the following script:

```
Script to execute:
python3 method.py -s 5.1 -in "<Cits_Dataset-Path>"

Example:
python3 method.py -s 5.1 -in "example_data/cits_dataset.csv"
```

doi		intext_citation.intent	intext_citation.sentiment	intext_citation.ret_mention
10.1002/jppr1998286443	...	TODO	TODO	TODO
10.1177/1362361398024006	...	TODO	TODO	TODO
...		...	...	...

step_5_1.csv

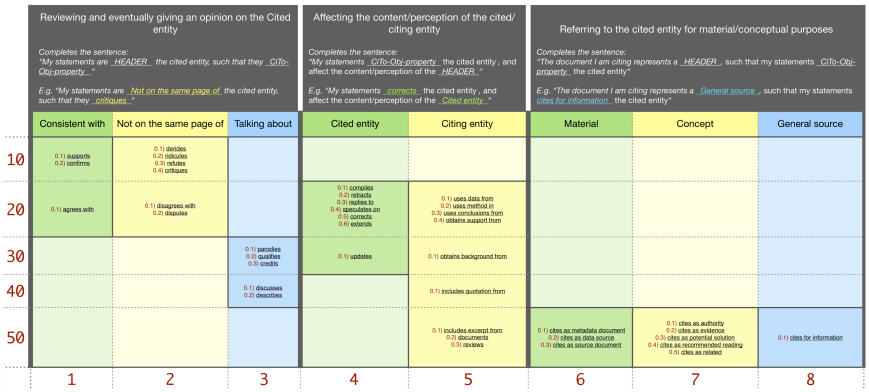
5.2 All the variables of this step are inferred by reading the intent citation context. This operation is made manually without the help of any script. To correctly replace the default initial "*todo*" values we will follow the below instructions.

The in-text citation intent ($intext\_citation.intent_{var}$):

This variable answers the question "Why the citing entity is citing the retracted article?", so we want to examine the intent/reason of the citation. The CiTo ontology, the Citation Typing Ontology (https://sparontologies.github.io/cito), is an ontology for the characterization of factual and rhetorical bibliographic citations. Although the CiTo ontology characterizes also the in-text citations lacking an explicit in-text citation pointer, we will not consider these variants. Instead, we perform the analysis on the in-text citations previously annotated (Step-4) which appear in the full-text with an in-text citation pointer.

On CiTo the citation intents are the object properties ($CiTo^{op}$), the $intext\_citation.intent_{var}$ is compiled using only one $CiTo^{op}$ value. Despite the fact that an in-text citation might refer to more than only one $CiTo^{op}$, our work restricts the decision to only one value. This decision has been taken in order to simplify the future elaborations on the annotated data and to limit the decision ambiguities.

To decide which $CiTo^{op}$ assign for the examined in-text citation context in case of multiple suitable $CiTo^{op}$ values, we have designed a CiTo decision-model. This model is based on a priority ranked strategy. The following figure shows a graphical representation of the model.



Considering a member m of $Cits\_Dataset$, our decision-model works as follow:

1. We read the $Cits\_Dataset[m, intext\_citation.context_{var}]$, and find the most suitable citation intent for it. The above model presents 3 macroblocks, we outline the suitable one/s to the analyzed in-text citation context taking a cue from the description, the usage, and the example of each block. Notice that the analyzed in-text citation context might be suitable for more than only one block.
2. Once we have chosen a suitable macroblock/s, we move toward a deepen selection of the suitable CiTo object property/s. At the end of this phase, we will have a set of citation intents obtained from

CiTo: $Intent\_set = \{x : x \in CiTo^{op}\}$.

3. In case we have chosen only one value: "*x*", then the value of $Cits\_Dataset[m, intext\_citation.intent_{var}]$ is "*x*". Otherwise (the $Intent\_set$ contains more than one $CiTo^{op}$) we take a decision based on a priority approach as described in the next point.
4. To calculate the priority of a $CiTo^{op}$ value: "*x*", we sum the corresponding y-axis and x-axis values, along with its cell inner value. The smaller a value, the more priority it has. For instance, $priority(\text{“}confirms\text{”}) = 11.2$ is higher than $priority(\text{“}describes\text{”}) = 43.2$. We will calculate the priority of each value "*x*" in the $Intent\_set$, and select the one with the higher priority value.

This analysis and discussion have much more details we decided to not mention in this methodology, since it goes out of the scope of this work. The documentation at https://ivanhb.github.io/intext-cits-ret-method/doc/cito_model.html, gives details and real use cases to better understand this method.

We might have more than only one $intext\_citation.intent_{var}$ value for the examined citing entity, since a record m in $Cits\_Dataset$ might have N number of in-text citations in its full-text, and therefore N different values in $Cits\_Dataset[m, intext\_citation.context_{var}]$. In this case, we must include the ";" segment as a separator between two different $intext\_citation.intent_{var}$ values.

The in-text citation sentiment ($intext\_citation.sentiment_{var}$):

After reading each $intext\_citation.context_{var}$ we will classify the citing entity and annotate the value of the $intext\_citation.sentiment_{var}$ with one of the following values:

- "*positive*": the retracted article was cited as a valid prior work, and its findings/conclusions could have been also used in the citing study.
- "*negative*": the citing study cites the retracted article but addresses its findings as inappropriate/valid.
- "*neutral*": the author cites the retracted article without including any judgment or personal opinion regarding its validity.

We might have more than only one $intext\_citation.sentiment_{var}$ value for the examined citing entity. In this case, we must include the ";" segment as a separator between two different $intext\_citation.sentiment_{var}$ values.

The in-text citation mentions the retraction ($intext\_citation.ret\_mention_{var}$):

We look at the value of $intext\_citation.context_{var}$ and check whether at least one of the in-text citation context values does explicitly mention the fact that the cited entity is retracted. Notice that at this stage we are not anymore interested in attributing singular characteristic to each inner value of the $intext\_citation.context_{var}$. We rather want to annotate with a "*yes*"/"*no*" value the $intext\_citation.ret\_mention_{var}$.

To make this annotation as much as possible coherent and less subject to language ambiguities, we decided to annotate a "*yes*" value, only if the word “retract” and its derivatives, are explicitly used while

addressing the cited entity in at least one of the in-text citations contexts in the $intext\_citation.context_{var}$.

doi		intext_citation.intent	intext_citation.sentiment	intext_citation.ret_mention
10.1002/jppr1998286443	...	disputes;; discusses;; disputes	neutral;; neutral;; neutral	no
10.1177/1362361398024006	...	discusses;; cites as evidence	neutral;; neutral	no
...		...	...	...

step_5_2.csv