EERQI

EUROPEAN EDUCATIONAL RESEARCH Quality Indicators

A project funded under the Socio-Economic Sciences and Humanities Theme (SSH)

SEVENTH FRAMEWORK PROGRAMME

Description of aMeasure

Measuring extrinsic quality indicators in educational research publications EERQI report

Daniel Stoye, Jenny Sieber

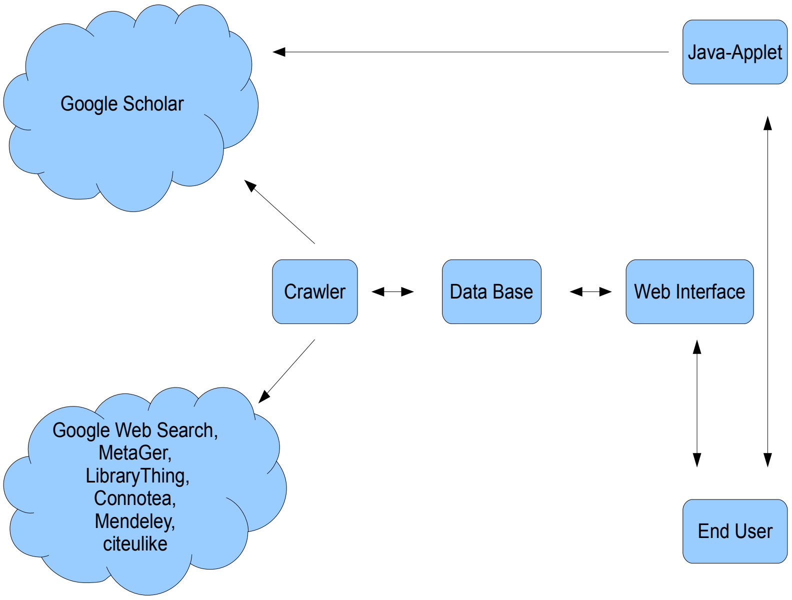
Content

1. Components....................2
2. Screen-Scraping of Google Scholar....................4
3. Filters....................5
4. Indices....................6
5. Limitations and Challenges....................7
5.1 Amount of results....................7
5.2 Foreign language characters....................7
5.3 Self citation....................8
6. Literature....................8

1. Components

aMeasure is a stack of tools and programs to measure extrinsic characteristics of research publications using Google Scholar[1], Google Web Search[2], MetaGer[3], LibraryThing[4], Connotea[5], Mendeley[6], and citeulike[7]. In the context of the EERQI project aMeasure will be used to collect information about extrinsic characteristics of educational research publications. It consists mainly of 4 parts:

- a crawler to gather all information from Google Scholar (GS), Google Web Search and the Social Network Services[8],
- a database[9] to store the gathered information,
- a client side application (JAVA-applet), and
- a web interface to present the results and the content of the database to end users.


Google Scholar
Java-Applet
Crawler
Data Base
Web Interface
Google Web Search, MetaGer, LibraryThing, Connotea, Mendeley, citeulike
End User

The main component of aMeasure is the crawler. For optimal work the crawler needs to be provided with author names. It has turned out that the major challenge in measuring extrinsic characteristics

[1] „Google Scholar provides a simple way to broadly search for scholarly literature. From one place, you can search across many disciplines and sources: articles, theses, books, abstracts and court opinions, from academic publishers, professional societies, online repositories, universities and other web sites.“ Retrieved from: http://scholar.google.com/intl/en/scholar/about.html 2010.10.04

[2] http://www.google.com/

[3] http://meta.rrzn.uni-hannover.de/

[4] http://www.librarything.com/

[5] http://www.connotea.org/

[6] http://www.mendeley.com/

[7] http://www.citeulike.org/

[8] Social Network Services mean applications like LibraryThing, connoetea etc.

[9] Not to be confused with the so-called “EERQI data base” in which the publisher's documents are stored.

of research publications is the reliable identification of author names in the Social Network Services, GS, Google Web Search, and MetaGer. We have therefore based our attempts on the findings presented by Derek Ruths and Faiyaz Al Zamal in the paper: “A Method for the Automated, Reliable Retrieval of Publication-Citation Records” published in 2010[10] . In this paper they present a series of filters to the results returned by an online publication search engine. One of these filters is a so-called name matching filter. Ruths and Zamal conducted several queries and retrieved “that when such a search is performed, the backend algorithm selects publications by applying a lenient filter to author names.” (Rutha, Zamal 2010, p. 3) They found that slight modifications of the authors name have a significant impact on the initial set of candidate publications returned by the search engine and therefore recommended to use the following query syntax: author:“the first name of the author the initials of the middle names the last name of the author. Using this syntax the crawler queries GS for the authors and all of their papers. This is done via Screen-Scraping[11].

In addition Google Web Search, MetaGer and the Social Network Services are queried to get information about the impact of each author's paper. The process of crawling is done on a central server located at HU Berlin and it is constantly running in the background.

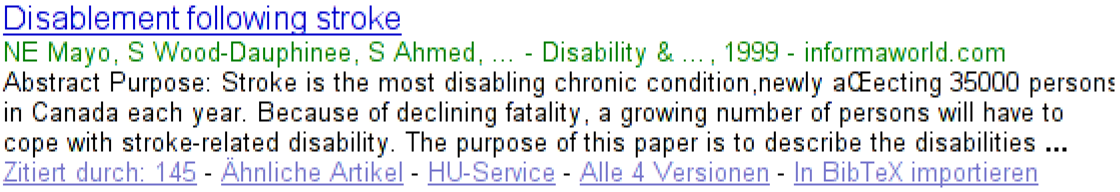
As Google has limited the number of requests to an unknown randomly selected amount per IP[12] per day the crawler is subjected to this limit too. If this limit is reached and a user intends to search for an author's name which has not been already stored in the central database, a Java-applet is querying GS instead of the crawler.

All data gathered, be it from GS or be it from the Social Network Services, are stored in a central Mysql database located on the EERQI server to enable various exports via the web interface.

2. Screen-Scraping of Google Scholar

GS is used to retrieve information about authors, their papers, and the citations of these papers. Due to the fact that Google does not provide an API[13] aMeasure is required to use a technology called Screen-Scraping[14]. The user normally issues a search request, so does aMeasure. This is achieved via URL-Parameters, for example: “http://scholar.google.com/scholar?as_q=author:"Ahmed, Sara"&hl=en” is doing the search: author:"Ahmed, Sara".

Afterwards the results page gets examined:



Disablement following stroke

NE Mayo, S Wood-Dauphinee, S Ahmed, ... - Disability & ..., 1999 - informaworld.com

Abstract Purpose: Stroke is the most disabling chronic condition,newly aŒecting 35000 persons in Canada each year. Because of declining fatality, a growing number of persons will have to cope with stroke-related disability. The purpose of this paper is to describe the disabilities ...

Zitiert durch: 145 - Ähnliche Artikel - HU-Service - Alle 4 Versionen - In BibTeX importieren

[10] http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0012133, retrieved: 2010.09.27

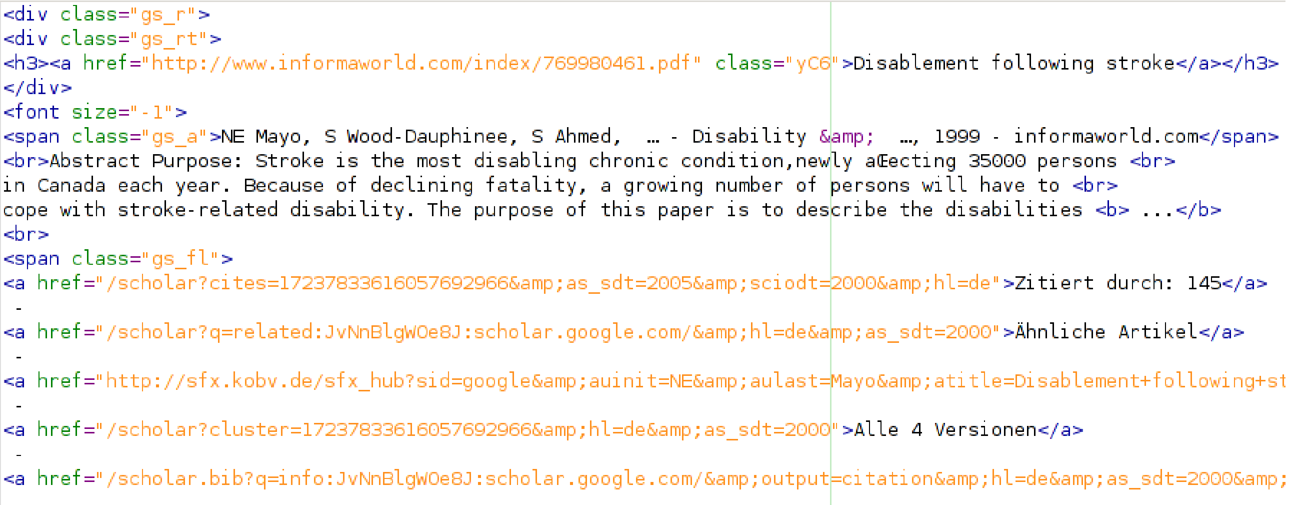
[11] For a detailed description of Screen-Scraping see section 2.

[12] Internet Protocol address

[13] http://en.wikipedia.org/wiki/Application_programming_interface, retrieved 2010.10.05

[14] http://en.wikipedia.org/wiki/Screen_Scraping, retrieved 2010.10.05

This excerpt is represented in the source code as:



```
<div class="gs_r">
<div class="gs_rt">
<h3><a href="http://www.informaworld.com/index/769980461.pdf" class="yC6">Disablement following stroke</a></h3>
</div>
<font size="-1">
<span class="gs_a">NE Mayo, S Wood-Dauphinee, S Ahmed,  … - Disability &amp;  …, 1999 - informaworld.com</span>
<br>Abstract Purpose: Stroke is the most disabling chronic condition,newly aŒecting 35000 persons <br>
in Canada each year. Because of declining fatality, a growing number of persons will have to <br>
cope with stroke-related disability. The purpose of this paper is to describe the disabilities <b> ...</b>
<br>
<span class="gs_fl">
<a href="/scholar?cites=17237833616057692966&amp;as_sdt=2005&amp;sciodt=2000&amp;hl=de">Zitiert durch: 145</a>
 -
<a href="/scholar?q=related:JvNnBlgW0e8J:scholar.google.com/&amp;hl=de&amp;as_sdt=2000">Ähnliche Artikel</a>
 -
<a href="http://sfx.kobv.de/sfx_hub?sid=google&amp;auinit=NE&amp;aulast=Mayo&amp;atitle=Disablement+following+st
 -
<a href="/scholar?cluster=17237833616057692966&amp;hl=de&amp;as_sdt=2000">Alle 4 Versionen</a>
 -
<a href="/scholar.bib?q=info:JvNnBlgW0e8J:scholar.google.com/&amp;output=citation&amp;hl=de&amp;as_sdt=2000&amp;
```

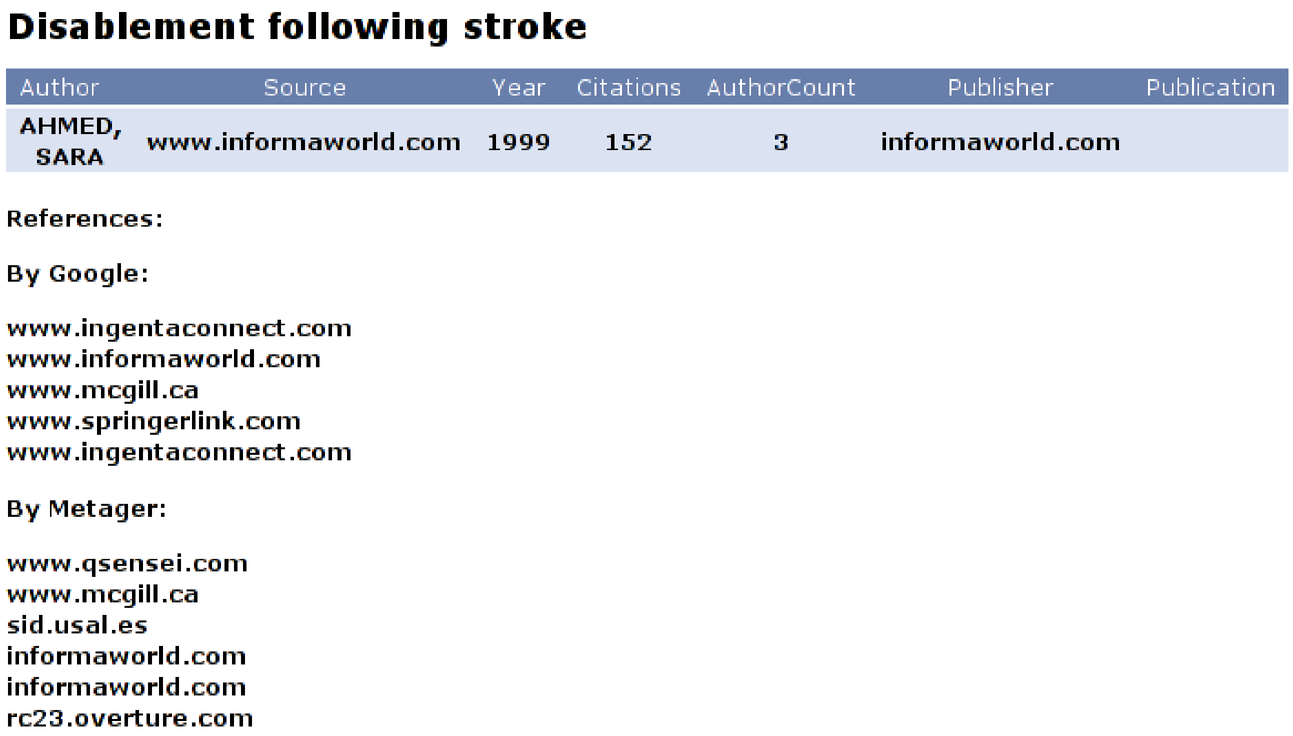
The source code is then parsed with the help of XML analysis. For example, every single result set (paper or book) is encapsulated by the HTML tags:

<div class="gs_r"><div class="gs_rt">

….result set data …

</div>
</div>

The page is split into these small parts, which are then further analyzed. For example, the document title is always encapsulated in <h3></h3> tags. With the help of these regularities, it is possible to identify every part of a result. In that way every part of the resulting item, be it the document name or be it the number of citations gets identified and stored in the database. This functionality is encapsulated into a Java-library to enable use in the crawler and the JAVA-applet which is working on the client side. The results are presented via the web interface at https://eerqi.hu-berlin.de/aMeasureWeb/ and are protected by project partner login.



Disablement following stroke

Author	Source	Year	Citations	AuthorCount	Publisher	Publication
AHMED, SARA	www.informaworld.com	1999	152	3	informaworld.com	

References:

By Google:

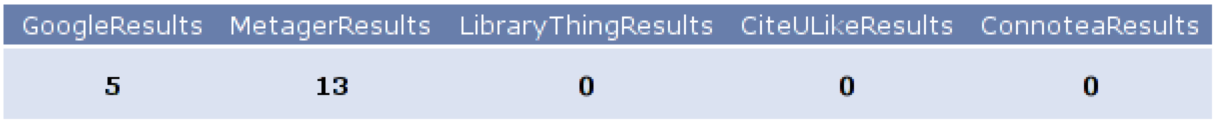
www.ingentaconnect.com
www.informaworld.com
www.mcgill.ca
www.springerlink.com
www.ingentaconnect.com

By Metager:

www.qsensei.com
www.mcgill.ca
sid.usal.es
informaworld.com
informaworld.com
rc23.overture.com

The same technology is used to query MetaGer and the Social Network Services. A more comfortable method is used for retrieving results from Google Web Search and Mendeley, which are providing APIs to their search engines.

These web search engines are queried with every single paper and the name of the author, for example: “Sahra Ahmed” + “Disablement following stroke”. The results are then presented via the web interface:



GoogleResults	MetagerResults	LibraryThingResults	CiteULikeResults	ConnoteaResults
5	13	0	0	0

3. Filters

We are aware of the fact that names are not unique. Relying on the “name filter” solely is not a suitable, sufficient criterion to discern the publications that belong to a given author. Since many individuals share the same last name, many more share the same first name. Taking this into account we integrated a second filter which ensures that the publications fall within the time span of an authors career. As we do not see how to get hold of each authors individual curriculum vitae we decided to limit the search results to the last 60 years arguing that an author is unlikely to start publishing before his/her 20th birthday and after his/her 80th year of life. Besides we take into account the results of the so-called “classifier”. This classifier contains a fingerprint of those word shingles (strings of defined length) which are typical for professional and relevant publications in educational research. The classifier can be asked via a API for a possibility if a given publication (identified by its URL) may be from educational research or not.

We also considered the idea of making the results more precise via a matching of author names and affiliations or places. We decided not to take into account the affiliations as we see a problem of standardization of e.g. institutions names and change of institutions names in general in the data sources we are using. We also decided to abandon the plan to make use of author-place matching even if the problem of name standardization and name changing seems to be not that drastic according to e.g. names of cities. But since we need the full coverage of an authors publications for the calculation of e.g. the h-index the limitation of an authors publications to just one place of his career seems to result in a distorted picture. Taking into account the rapid movement of especially young researchers we would run the risk of losing a large amount of publications. Searches for e.g. “Stefan Gradmann” + “Berlin” resulted in much fewer hits than searching for “Stefan Gradmann” + “Hamburg”, though we knew from the curriculum vitae that it is one and the same person in both cases.

4. Indices

The following extrinsic characteristics can be retrieved and calculated from GS using aMeasure:

- Number of papers per author.
- Number of citations per author.
- Year – first year of retrieved publication until last year of retrieved publication.
- Citations per year.
- Citations per paper.
- The h-index provides a single-number metric of an academic's impact. A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np−h) papers have at most h citations each. The h-index is calculated based on the full list of an authors output and the obtained citations. The *h*-index is robust in the sense that it is insensitive to a set of uncited or lowly cited papers but also it is insensitive to one or several outstandingly highly cited papers. This last aspect can be considered as a drawback and we therefor take into account the g-index.
- The g-index is an improvement of the h-index. It gives more weight to highly-cited articles. (Egghe 2006)
- The e-index is aiming to differentiate between scientists with similar h-indices but different citation patterns. (Zang 2009)

The following extrinsic characteristics can be retrieved and calculated from Google Web Search and MetaGer using aMeasure:

- Google Web Search hits matching the authors name.
- MetaGer hits matching the authors name.

The following extrinsic characteristics can be retrieved and calculated from Social Network Services using aMeasure:

- citulike hits matching the author's name and the articles title.
- LibraryThing hits matching the author's name and the articles title.
- Connotea hits matching the author's name and the articles title.
- Mendeley hits matching the author's name and the articles title.

5. Limitations and Challenges

5.1 Amount of results

GS and Google Web Search have the unpleasant habit to present an estimated result count only, due to that every user and every API request is not able to see or get more than the first 1000 results for a specific search request. In terms of Google Web Search the company has shut down their old XML-API which enabled users to get very close to these 1000 results. Currently the Google-AJAX-API is limited to 64 search hits. If the Google Web Search reaches 64 hits, we are using "Screen Scraping" of Google Web Search to get the full list of results.

5.2 Foreign language characters

Google also has some limitations regarding umlauts and accents:
The french author François Hochepied for example generates different results when written with or without "ç". The query
author:"François Hochepied" is resulting in less results than:
author:"Francois Hochepied"
Also,
author:"Malet, Régis" is resulting in less results than:
author:"Malet, Regis"

The troubles caused by German umlauts are much more problematic :
author:"Norbert Bläsing" and
author:"Norbert Blaesing" are generating equal results, but
author:"Norbert Blasing" leads to no results.

What we need is a single, unique method to identify authors as it is the critical step in making it possible to automatically track all the contributions that a researcher has made. This problem is very well known. In 2006 Elsevier launched its service "Scopus author identifier". The author identifier assigns a unique number to the authors who have published articles in journals covered by Scopus. An algorithm distinguishes those with similar or identical names on the basis of their affiliations, publication history, subject areas and co-authors.(Qiu, 2008) Scopus excludes records from the process that lack sufficient data to determine a match. Once clearly identified, authors receive a unique identifier number. In 2007 *CrossRef* invited a number of people to discuss unique identifiers for researchers[15] In 2008 Thomson Reuters launched ResearcherID. ResearcherID tries to solve exactly the above illustrated problem. In the PLoS Comp Biol article Bourne and Fink argue that one solution to this difficulty is *OpenID*[16]. OpenID is a standard. "That means that an identity can be hosted by a range of services and people can choose between them based on the service provided, personal philosophy, or any other reason. The central idea is that you have a single identity which you can use to sign on to a wide range of sites. There are two major problems with OpenID. The first is that it is poorly supported by big players such as Google and Yahoo. Google and Yahoo will let you use your account with them as an OpenID but they don't accept other OpenID providers. More importantly, people just don't seem to get OpenID"[17]

[15] http://www.crossref.org/CrossTech/2007/02/crossref_author_id_meeting.html

[16] http://openid.net/

[17] http://blog.openwetware.org/scienceintheopen/2009/01/20/a-specialist-openid-service-to-provide-unique-researcher-ids/

This state of our knowledge clearly isn't satisfactory and requires additional work in the future. A first attempt to overcome this unsatisfactory situation is in process. A heuristics listing all possible combinations of umlauts and accents in a given name is going to be developed. For example:

Malet, Régis
Malet, Regís
Malet, Régís
Malet, Régìs
Malet, Regis

Malet, Règis
Malet, Regìs
Malet, Règìs
Malet, Règís

This heuristics will be the basis for gathering information from GS and the other resources. All hits matching the queries will be listed.

5.3 Self citation

Currently aMeasure is filtering self citations with the help of GS. By using GS it is possible to search within all citations a paper has received. By subtracting all citations where the author of the original paper is also the author or co-author of the citing paper from the total amount of citations the paper has received we can filter out self citations. This technique prevents us from analyzing all citations manually, which would involve many queries to GS and would reduce the amount of papers and authors we are able to analyze per day. As some authors published a lot of papers which obtained many citations, and as there is a daily limit GS sets per user or IP per day this solution seems to be the most comfortable one in terms of returning hits in a reasonable time. From our point of view tools like CleanPoP[18] do not seem to take this into account or present just a limited number of results concealing the illustrated problem of limited requests to Google. Besides one further drawback of CleanPoP is the necessity to manually select author names and possible duplicates. This means that every single citing paper needs to be analyzed.

6. Literature

Al Zamal, F., with Ruths, 2010. A Method for the Automated, Reliable Retrieval of Publication-Citation Records. *PLoS One*, 5(8). Available at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0012133.

Bourne, P.E. & Fink, J.L., 2008. I Am Not a Scientist, I Am a Number. *PLoS Comput Biol*, 4(12).

Egghe, L., 2006. Theory and practise of the g-index. *Scientometrics*, 69(1), p.131–152. Available at: http://www.springerlink.com/content/4119257t25h0852w/fulltext.pdf.

Qiu, J., 2008. Scientific publishing: Identity crisis. *Nature*, 451, pp.766-767. Available at: http://www.nature.com/news/2008/080213/full/451766a.html.

Wolinsky, H., 2008. What’s in a name? *EMBO reports* , 9, pp.1171 - 117. Available at: http://www.nature.com/embor/journal/v9/n12/full/embor2008217.html.

Zhang, C.-T., 2009. The e-Index, Complementing the h-Index for Excess Citations. *PLoS ONE*, 4(5). Available at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0005429.

[18] http://cleanpop.ifris.net/