

## **Assessing the Impact of Legal Publications Building a Legal Citation Index using Automatic Reference Analysis**

Kees (C.) van Noortwijk

### **Abstract**

In many fields of science, 'impact factors' are assigned to articles, authors and magazines as a matter of routine. Databases containing abstracts and citations of scientific publications are available to facilitate this. However, in the Netherlands this is not the case for publications in the field of Law. For several reasons, these publications are not generally 'ranked' according to their importance and/or 'impact' and citations from these publications are not compiled and/or processed systematically.

The need to change this situation is generally felt. One reason is that important publications can be more easily found in the ever increasing amount of digital data available to researchers when information relating to their impact is available. Another reason is that universities and research institutes need tools to assess the work of their employees. But the construction of a legal citation index from scratch is no small task.

The project described in this paper aims at contributing to a more effective impact measurement of Dutch legal publications – for instance, articles in journals – by facilitating the automatic collection of citation data from publications. This is accomplished by making use of so-called content integration technology.

Using the collected citation data, it is possible to assess the 'impact' of the publications involved much more reliably.

### **1. The 'impact' of a scientific publication**

Scientists in every field produce a continuous stream of publications. In many cases these constitute their most important channel for communicating the results of their scientific work. A scientific publication often marks the completion of (a particular part of) a research project and makes details and results from this project available to interested parties.

Not every publication, even if its contents do conform to the general standards for

scientific work<sup>[1]</sup>, is of the same significance. This could be related to the importance and/or the quality of its contents, but also to the audience it finds. At the same time, the assessment of the results of research projects – which often primarily are made public in the form of scientific publications – these days is seen as essential. One of the reasons is that the means devoted to scientific research projects need to be justified, specifically towards those that provide funding. For any project that is funded 'externally' (i.e. by a party outside a university or research institute), such an assessment in fact finalises a sequence that starts when a researcher requests funding via a research proposal. Furthermore, it can be of importance for the professional evaluation of the work of researchers, for instance in the context of a tenure track. A correct evaluation of not only the number and the type, but also of the quality of scientific publications, therefore seems of importance to many parties.

Specifically in the field of law, in The Netherlands a true qualitative evaluation of legal scientific publications currently is not easy. In fact, the main criterion that is used to that extent at this time is the presence or absence of any form of editorial review, be it by peers or by a board of editors, which serves to judge papers prior to publication in a particular journal. Based on that criterion, (papers in) such journals are generally trusted to be of higher quality than those published in unreviewed periodicals. As exceptions to that general rule are usually quite easy to find – important papers sometimes being published in an unreviewed source, and papers of questionable quality nevertheless passing review – other, more objective and substantive ways to estimate the quality of scientific output are definitely desirable.

## 2. Citation indexing

The term 'citation' is commonly used to indicate references to published or unpublished sources, found in other publications. In a scientific publication, such references are used to acknowledge the work of others and are generally indispensable to uphold intellectual honesty (or avoiding plagiarism), to attribute prior or unoriginal work to the correct sources, to allow the reader to verify whether the referenced material indeed supports the author's argument and to help the reader assess the validity of the material the author has used.<sup>[2]</sup> The idea to count and use citations of a certain publication or author to establish the scope of the audience reached, and connected with that, the 'impact' of the publication or (the work of) the author, already has a long history. Early examples with respect to scientific literature date back to the second half of the nineteenth century.<sup>[3]</sup> The citation index known as 'Web of Science'<sup>[4]</sup>, originally named 'Science Citation Index' (SCI) was initiated by the Institute of Scientific Information (ISI) in 1964<sup>[5]</sup> and is one of the most extended citation indexes of scientific work. Its contents form the basis for the calculation of 'Journal Impact Factors', a measure reflecting the yearly average number of citations to recent articles from a particular journal. In general, methods like these are part of the field of *bibliometrics*.<sup>[6]</sup>

Specifically in social sciences, the analysis of citations to a publication is recognised as a valid method for assessing the 'impact' of that particular publication, although the method used for that analysis is certainly subject of ongoing discussion.<sup>[7]</sup>

Quantitative data obtained from citation analysis should be used with some caution. As has been pointed out, huge differences may exist between the number of citations of different papers in the same journal, and even of the same author. That is one of the

reasons the use of mathematical averages to calculate impact factors for journals has been widely criticised.[\[8\]](#)

Still, gathering citation data at least provides us with some basic notion of which publications attract most attention, which subject areas are currently under discussion and which authors (scholars) are involved in those discussions. That seems to be useful and important information for anyone evaluating the activities within a certain field, and the contributions specific authors make to that. After all, as Moed (2005) states, “citations of references can be conceived as social acts of members of a scholarly community”.[\[9\]](#) Let’s assume, therefore, that notwithstanding the reservations one might have about journal impact factors, estimating the impact (based on citations) of an individual article could be a useful exercise. Of course, such a measure for an individual article would always have to be compared to that of other articles. Issues like differences in size would have to be taken into account. But given that, obtaining base data on citations of an individual article could definitely help to better estimate the amount to which that article ‘gained traction’, which could be an indicator of its impact in the field.[\[10\]](#)

### 3. Project goals and methodology

#### 3.1 Assessing the impact of Dutch legal publications

In this paragraph, what has been said about citation indexing is applied for Dutch scholarly legal publication. For these publications, the evaluation of their ‘impact’ – for instance, established from the number of times it is referred to (or ‘cited’) in other publications and by other authors – is not a standard assessment criterion, whereas that certainly is the case for publications in fields like physics or social sciences. An important reason for that is the absence of a useable registration system for citations, a ‘citation index’, in which these publications are represented. Consequently, it is in fact barely possible to find out if an article in any of the Dutch legal periodicals is cited often, or not at all, without performing a lengthy analysis ‘by hand’.

To gain insight in the relative importance of articles in these publications, basic information about the number of citations of these articles would be beneficial. Unfortunately, the amount of data currently available for acquiring such insight is very limited. Citation indexes such as Web of Science and Scopus contain practically no information with respect to Dutch language legal periodicals.[\[11\]](#) Of course, that does not mean that Dutch lawyers have no clue whatsoever as to the importance of a certain publication. They rely on the editorial board of the journals that cover their area of law to guard the overall quality of what is published in those journals. But more detailed information on the effect that a publication achieves in the field could definitely be useful. Law is usually not seen as a ‘hard science’, but it is definitely a field in which authority plays an important role. And this authority could certainly be connected with the audience that is reached by publications of a particular legal author. An analysis as will be outlined here could possibly provide insight with respect to that.

Building a separate citation index for a specific category of publications is not uncontroversial. It has been argued that a comprehensive index, covering as wide a selection of publications as possible, forms the best prerequisite for recognising a representative proportion of all citations of a particular publication.[\[12\]](#) After all, the larger the database, the higher the probability that all (or at least, a high number of) citations

can be found in it. Furthermore, Dutch law functions within a European – or even global – system of international legislation, which might lead to foreign authors showing an interest in publications about Dutch law, and possibly to citing these. This citation in a foreign publication would probably be missed in a Dutch legal citation index.

Although these are relevant arguments in favour of an integrated international index, which I think should indeed be an ultimate goal, there are also reasons not to wait for that, and follow a different approach in this particular case. One of these reasons is that the vast majority of scientific publications on Dutch legal issues appears in journals in the Dutch language, most of which are currently not covered in any citation index.<sup>[13]</sup> When we want to estimate the impact of these publications, it makes sense to take into consideration the citations in Dutch language legal sources, which will probably by far outweigh possible citations in foreign sources.

Specifically when the main aim is to 'weigh' individual publications (and, optionally, the authors, by combining the data of their separate publications), I think that in this case the use of citation data based on just the Dutch language subset of legal publications is useful and justified. The data gathered for that could possibly also form the basis for adding these publications to international citation indexes in the future, which would definitely be an added benefit.

### 3.2 Sources of citation data

What data would be needed to newly construct a citation index for a particular field for which no citation data are available yet? A preliminary list of the most important data comprises the following.

- Bibliographic data of all publications involved (further referred to as 'identifying data'), including
  - the title of the publication;
  - its author(s);
  - the date (or at least the year) of publication;
  - if the publication appeared in a journal, is part of a book or bundle or some other compilation work: the title of that journal, book or other work (the 'main publication') and its author(s) and/or editor(s);
  - (optionally) the size of the publication, or the number of pages;
  - other information useful for identifying the publication (for instance: edition number of the journal, serial number of the publication in that edition or in the year of publication, etc.).
- For each publication involved: a list of identifying data of all other publications that are cited in that publication. The list should contain the information needed to identify the respective publications unambiguously.

To satisfy this last demand, the information about the publications cited should either consist of a combination of title, name(s) of author(s) and year of publication, or of the name of the main publication, edition number and publication serial number, or of any

other combination of the data mentioned above suitable to uniquely identify the publication that is cited. Usually, this information about publications that are cited can be obtained from the (citing) publications themselves. If these contain references to other work, the references should contain all information needed to identify and find that other work. And indeed, we are usually capable of doing so.

Building a citation index 'by hand' would therefore in itself not be a complicated task. We could start with a series of recent publications, store the identifying data for each of them together with the identifying data for each of the publications cited by them and repeat the process with these publications as well. The data collected in this way will eventually make it possible to list, for each publication,

- all other publications that the publication refers to (or which it 'cites'), as well as
- all other publications that refer to the publication (or 'cite the publication').

The other publications that refer to this publication, are commonly named the 'citations' of the publication and this terminology will be used here, too

Although not complicated, it would of course not be very practical to collect citation data 'by hand'. The number of publications involved can quickly get out of hand. The number of journal articles only, appearing in Dutch legal periodicals in the year 2017, is already higher than 30.000.<sup>[14]</sup> When starting from scratch, it could take years to process (i.e. store the data for) this number of articles. And during that time, equal numbers of new publications would already have appeared. Therefore, the only practical way to gather citation data covering a somewhat more extended period of time, would be to automate the process.

Collecting citation data automatically is possible when the publications involved (those referring, as well as those referred to) are available digitally. For recent journal articles, that is practically always the case these days. In fact, for most periodicals digital archives are available that date back at least ten years. That means that, in principle, it would be possible to find and fully process all references in an article if the sources that the article refers to are up to ten years old. For all these sources, the 'citations' (the list of articles referring to them) could at the same time be established.

One could ask why it is desirable that *all* sources involved (not only the referring article, but also the ones referred to) should be available digitally. Why isn't it enough that the referring article can be processed digitally, with all citations found in it just stored (and later combined)? The reason is, that it is vital that the publication referred to is *recognised positively*, in order to correctly gather all citations for that publication. The information found in references, however, is often far from complete. The title might be misspelled, names of authors might be omitted, etc. Therefore, it is challenging to collect citations just by registering the data in references. The process can be completed much more reliably when the article referred to, with all its metadata, is available digitally too.

### 3.3 Content integration

A development I would like to describe at this time, is that of so-called *content integration systems*. These online systems, which have been available in The Netherlands for about ten years now<sup>[15]</sup>, can be of particular interest for the subject of these studies. The reason is that by means of content integration systems, extended sets of legal publications can not only be made available through one user interface, but can also be linked

together, in the sense that when a certain publication refers to another one (for instance, in a footnote), this reference can be converted to a working hyperlink. Especially this last possibility is of importance here.

With respect to these content integration systems, the 'legal IT landscape' is quite unique in the Netherlands. Like everywhere else, the main legal sources are produced by a few large and a series of small publishing houses. Large publishers include Wolters Kluwer, Sdu (part of ELS now) and Reed Business, smaller ones include Boom Juridisch, Paris, Den Hollander, Delex, and over ten others. Each of them offers sources digitally, often using a proprietary system for online disclosure. Kluwer, for instance, offers the Navigator system, which gives access to practically all of their legal publications. The system makes it possible to search these sources by means of a full text search interface and to browse through their contents via digital tables of contents. Although such one-publisher solutions usually work satisfactorily, they have one main flaw: they cannot access any content from other publishers. And because they cannot access it, they are also unable to link (via hyperlinks), from the publications they disclose, to any other content but their own.

Many law firms in The Netherlands use content from at least five of these publishers. Furthermore, they also use open access legal content, such as the case law published by the Judiciary and the legislation published by the Dutch government. Using all these sources would normally require them to adopt and use at least six or seven different information retrieval systems, each with its own user interface and its own possibilities and limitations. Around ten years ago, when the top 50 law firms started to switch from paper to digital information sources, resistance against this situation grew. Several IT companies started offering 'integrated retrieval services', which at first only gave access to open access sources and to sources from small publishing houses (which were only happy to leave this service to a third party, so they did not have to build and maintain their own retrieval systems). But when law firms saw the benefits of such 'integrated solutions', they put pressure on the larger publishers to also provide their content for integrated access. And because these firms were very important customers to the publishers, their efforts were effective eventually.

This led to a situation in which, at present, practically all digital legal publications are not only available through the systems of their respective publishers, but also via several content integration systems. Two of these systems currently dominate the market: Rechtsorde and Legal Intelligence. Although both these systems were at a certain point obtained by a large publisher after having started as independent enterprises, they can still be used to access content from all other publishers as well; they have gathered enough market power to guarantee and maintain this service.

The reason why all this is mentioned here, is that content integration provides essential facilities for the gathering and compilation of citation data, which in turn can be used to build a citation index. We could call this an extremely useful side effect of these systems. The most important characteristic that makes this possible, is the fact that content integration systems disclose – to the extent that this is possible – *all* (or at least the majority of the) legal source material, via one search system. Because all separate documents – articles, chapters from books, papers, even news items – are present and are known to the system, it can recognise, with far greater precision than would otherwise be possible, the citations present in each of these documents. That is, if these references indeed refer to one of the other documents in the system's database. In some cases, these references in fact already are provided in the form of hyperlinks, linked to the actual



document they refer to. That is mainly the case when the reference is to a source of the same publisher, or to an open access source. Such linked references are available in the document as added metadata.

### 3.4 Link recognition

The real challenge, however, is to convert *every* reference in a document into such a hyperlink, capable of opening the document that is cited directly. To achieve that, at least one of the content integration systems – namely Rechtsorde – performs ‘link recognition’ on the complete text of every document. When in that text a sequence of characters is spotted that conforms to the format of a reference, the system tries to ‘solve’ that reference by comparing it to (the characteristics of) all known documents. It is easy to understand that this process works best when the number of documents available in the system (those are the ones that are ‘known’) is as high as possible.

The process of recognising references to publications in the text of a document is described by Van Opijnen (2014, p. 371-393). As he explains, for this recognition process so-called *regular expressions* are commonly used. These can be derived from the names (or abbreviations for names) of publications, and work as a filter or template capable of identifying every piece of text that constitutes a valid reference to that publication.<sup>[16]</sup> With carefully created regular expressions for each publication, it should be possible to recognise every reference to documents (articles, chapters) from that publication, unless the reference is incomplete or incorrectly formed.

In a situation where scientific publications are issued by different publishers and each of these publishers only provides linking metadata for citations of publications issued by that same publisher, link recognition is in fact the only way to interconnect *all* relating publications automatically. But apart from that, at the same time it can also be used as a tool to identify and validate the citations present in every publication. During the link recognition process, the basic data necessary for creating a citation index – namely the list of all citations present in each document – can be generated completely automatically. A possible application of that technology will be described in the next paragraph.

## 4. Processing citation data from digital publications

### 4.1 Data from the link recognition process

As described in the previous paragraph, the link recognition technology used in certain information disclosure systems can provide the basic data from which a citation index could be constructed. What exactly are these data?

When the author of a (in this case legal) scientific publication cites other publications, this is usually done by mentioning the author of that publication, its title, the year of publication and the source in which it was published. Other data, such as the name and location of the publisher, might also be given, but are usually not essential for identifying the publication. The source of the publication might be indicated by giving its full name, but also by giving a known abbreviation for that name. For example: a well known periodical such as the ‘Nederlands Juristenblad’ can reliably be abbreviated as ‘NJB’. Recommendations exist as to the best way to indicate existing legal sources.<sup>[17]</sup>

However, these recommendations unfortunately are not followed by many authors. Therefore, a link recognition algorithm should take into account several variants of citations of certain sources.

The data that could be recorded during the link recognition process of a new publication, would consist of the following sets of data, one for each citation identified:

- [publication A]
  - of [author A']
  - in [source A'']
- contains a reference to [publication B]
  - of [author B']
  - in [source B'']

These data can be stored in a relational database, adding unique (key) identifiers for each publication, author and source. By querying this database, after the link recognition process is finished, not only can basic data for construction of a citation index be obtained, but also potentially useful statistics, such as quantitative data about the average number of recognised citations per author and per source.

#### 4.2 The reliability of link recognition

While performing link recognition, the possibility always exists that a certain variant of a reference to a source is not recognised correctly. This would lead to that particular citation staying unnoticed by the system. An important question is therefore: how reliable is the link recognition process, as it is currently applied in practice? To answer that question, an analysis of some results of link recognition in the Rechtsorde retrieval system – which as stated earlier is known to apply this technology to legal publications – was made.

This analysis was performed in the following way. First, a sample containing some suitable journals was defined. The only criteria used here were that the journals publish articles in which citations could be present, and that the journals were available digitally via the Rechtsorde system. The first criterion was necessary, as numerous available journals solely publish case law (with or without annotation), which usually does not contain citations of other publications. Thirteen different journal titles were selected at this time (with the possibility to include more titles later). From each of these titles, one or two articles from past issues were randomly selected. Table 1 lists the journal titles and the article numbers selected.



| Journal title:                                              | Article / Issue | Publisher      |
|-------------------------------------------------------------|-----------------|----------------|
| Tijdschrift Arbeidsrecht                                    | 2015/47         | Kluwer         |
| Aansprakelijkheid, Verzekering en Schade (AV&S)             | 2013/22         | Kluwer         |
| Bouwrecht (BR)                                              | 2014/99         | Kluwer         |
| Computerrecht                                               | 2009/36         | Kluwer         |
| Computerrecht                                               | 2014/72         | Kluwer         |
| Delikt en Delinkwent (DD)                                   | 2006/78         | Kluwer         |
| Delikt en Delinkwent (DD)                                   | 2010/60         | Kluwer         |
| Expertise en Recht (EeR)                                    | 2013/04, p.141  | Paris          |
| Tijdschrift voor Familie- en Jeugdrecht (FJR)               | 2012/105        | Kluwer         |
| Nederlands Tijdschrift voor Bestuursrecht (NTB)             | 2014/30         | Kluwer         |
| Regelmaat                                                   | 2015/04, p.279  | Boom Juridisch |
| Tijdschrift voor Pensioenvraagstukken (TPV)                 | 2014/48         | Kluwer         |
| Tijdschrift voor Insolventierecht (TvI)                     | 2015/50         | Kluwer         |
| Tijdschrift voor Procespraktijk (TvPP)                      | 2012/04, p.91   | Paris          |
| Weekblad voor Privaatrecht, Notariaat en Registratie (WPNR) | 2010/6827       | Sdu            |

**Table 1 - Journal titles and articles to analyse**

As all of the articles in Table 1 are available via the Rechtsorde system, each of them was known to have been processed by Rechtsorde (prior to making the complete journal available). As part of this processing, the link recognition process should have been applied to the articles. The question was, however, in what way the results of that could be made visible.

The answer to this proved to be relatively simple. Whenever a reference in an article is recognised by this particular retrieval system, a hyperlink to the document cited is added to a list of 'Related publications'.<sup>[18]</sup> Therefore, the following procedure was performed:

1. The articles from Table 1 were retrieved one by one, using the standard search interface, and were opened.
2. From the text of the article, all citations (often present in footnotes or endnotes) were copied manually to a separate file.
3. From these copied citations, only those that were *not* presented in the form of hyperlinks in the article itself, were marked. Reason for this: citations with hyperlinks are probably accompanied by metadata, already specifying the correct

document cited, and are therefore not the product of the link recognition process under test here.

4. Next, the list of 'Related publications' for that article was opened. It was checked if each of the marked references from step 3 could be found in the list. The outcome of this test for each separate citation was registered.

The result of this procedure was a list of all citations from the articles chosen, filtered to only contain those citations that were *not* accompanied by metadata, with for each of these citations an indication if the link recognition process had correctly found the right document for that citation or not. Table 2 shows some citations that were *not* recognised correctly. In this list, only those citations are mentioned that contain enough elements that could in principle have been recognised.

| Citation not recognised (exact copy):                                       | In publication              |
|-----------------------------------------------------------------------------|-----------------------------|
| Justitiële Verkenningen. Mensenhandel en -smokkel, 1996, p. 46.             | Delikt en Delinkwent 2010/8 |
| Proces 2009, 5.                                                             | Delikt en Delinkwent 2010/8 |
| Gst. 2014/67                                                                | Bouwrecht 2014/99           |
| O&A 2013/32                                                                 | Bouwrecht 2014/99           |
| Tijdschrift voor Gezondheidsrecht 2012 (36)1, p. 16.                        | AV&S 2013/22                |
| TCR 2011, nr. 2, p. 48.                                                     | AV&S 2013/22                |
| Tijdschrift Financiering, zekerheden en insolventiepraktijk 2014/3, p. 106. | TVI 2015/50                 |
| M. Vetter, Insolventierecht, dertiende druk, Deventer: Kluwer 2014          | TVI 2015/50                 |
| Trema (36) 2013-3, p. 89-96                                                 | EeR 2013/04, p.141          |

**Table 2 - Partial list of citations not recognised**

In most cases, the reason why the citation had not been recognised was not difficult to determine. In some cases, an incorrect abbreviation or full name for a magazine had been used, or the title of the article referred to was placed between the journal name and the year of publication. Furthermore, the various ways in which page numbers were used in the citations often seem to be outside the scope of the link recognition mechanism and therefore are not processed correctly. These recognition errors were passed on to the manufacturer of the Rechtsorde system, to promote future improvements.

Furthermore, there proved to be quite a few citations to sources that were not available in Rechtsorde. These were mainly citations of book publications, foreign sources and non-legal publications. Such citations could not be recognised because the system had no knowledge about the respective sources. This is exactly where the results from this content integration approach fall short of results that could have been achieved when all publication involved (citing and cited) would have been part of a comprehensive, global citation index such as Web of Science. The question is, however: are the results achieved still useful?

The data with respect to the number of citations present in the articles, and the number of these being recognised correctly, are shown in Table 3. The table contains three columns of numbers. The first column specifies the number of citations present in the respective article. This number excludes any citations in the form of hyperlinks (for which recognition is irrelevant) and also any citations of (unannotated) case law. The second column contains the number (subset) of these citations that specify a publication actually available in the Rechtsorde system. And finally, the third column specifies the number of those citations to known material that were actually recognised correctly.

| Journal/article                               | Citations present | Of which in RO | Recognised |
|-----------------------------------------------|-------------------|----------------|------------|
| Arbeidsrecht 2015/47                          | 19                | 11             | 3          |
| AV&S 2013/22                                  | 6                 | 5              | 0          |
| BR 2014/99                                    | 56                | 47             | 32         |
| Computerrecht 2009/36                         | 10                | 0              | 0          |
| Computerrecht 2014/72                         | 10                | 8              | 7          |
| DD 2006/78                                    | 13                | 6              | 0          |
| DD 2010/60                                    | 11                | 8              | 4          |
| EeR 2013/04, p.141                            | 24                | 2              | 1          |
| FJR 2012/105                                  | 15                | 3              | 3          |
| NTB 2014/30                                   | 19                | 4              | 2          |
| Regelmaat 2015/04, p.279                      | 24                | 15             | 5          |
| TPV 2014/48                                   | 24                | 12             | 2          |
| TvI 2015/50                                   | 23                | 13             | 11         |
| TvPP 2012/04, p.91                            | 9                 | 4              | 2          |
| WPNR 2010/6827                                | 16                | 13             | 7          |
| Total:                                        | 279               | 151            | 79         |
| Percentage found of present in RO system      |                   |                | 52.30%     |
| Percentage found of total number of citations |                   |                | 28.30%     |

**Table 3 - Results of link recognition analysis**

From a total of 279 citations present in these 15 articles, 151 citations were to sources present in and therefore known to the Rechtsorde system. Of these 151, 79 were correctly recognised, the relevant documents showing up in the list of 'related documents'. This means that from the citations that could actually have been recognised, 52.3% were recognised correctly. Of all citations present (excluding citations to unannotated case law etc.), 28.3% were recognised.

The results show that link recognition technology is still far from perfect. Only about half of all citations that could in principle have been recognised, were actually processed. Still,

this result can definitely be seen as promising. It has in fact improved considerably in the past few years. Therefore, the percentages of correctly recognised citations given here should be considered no more than a snapshot result, as the link recognition process is known to be constantly evaluated and adjusted. The conclusion from this experiment is therefore that link recognition can indeed be effective, even though the technology is not perfect yet.

Based on these results, expectations could be that of all available reference info in all publications available digitally, around half can already be saved automatically as citation data. This percentage will probably grow even higher in the future, as recognition technology is gradually improved.

### 4.3 Possible implementations

The automatic processing of citations in legal literature, as tested in the way that was described in the previous paragraph, could be implemented in several ways.

The first, and probably simplest solution would be when extra functionality could be built into the integrated retrieval systems, such as the one used here. The result could be a system that solves references automatically, and processes these to automatically produce all available citation ('cited by') information with every publication. This would be a different function from the 'Related publications' function that was described earlier, because in this case the citation information would be attached to the publication referred to. A drawback of that approach would be that it would not be easy to obtain compiled citation information for all publications in the database. The only way to achieve that would be to consult the citation data for every separate document, which would definitely not be practical.

A better solution would therefore be to develop a separate tool, capable of gathering the citation data from the content integration system's database. The advantage of that would be that the tool could be run periodically, updating citation data whenever new publications become available. As described earlier, a flexible solution for this would be to have this tool store the citation data in a new relational database. That database could then be queried, to obtain data such as

- A list of all publications that contain citations, with the number of citations per publication;
- A list of all publications which are cited (in another publication), with the number of citations per publication;
- A list of citations for any given publication;
- A list of all authors, with the total number of citations that each of them has used in his/her publications;
- A list of all authors, with the total number of citations of each of their publications;
- The minimum, maximum and average number of citations of publications from a particular source;
- The minimum, maximum and average number of citations of (publications of) a particular author;

etc.

The output of these queries would deliver all (base) information needed for the execution of an impact analysis for a certain article or author. Furthermore, a citation index could be generated from this, although it might be necessary to store some extra bibliographic data in the database in order to be able to show these data with every publication.

## Conclusions

In this paper, it is argued that it is possible to gain insight into the (objective) significance of legal publications, even if no citation data for such publications are readily available. By using existing mechanisms for the integrated disclosure and the automatic linking of digital sources, citation data can be compiled at least partly automatically. An (admittedly limited) test showed that around one third to half of the available references in a certain publication can already be processed automatically.

This means that, even though improvements are still desirable, estimations with respect to the 'impact' of publications, based on for instance citation data, are already a possibility, even in fields of science for which no citation data are available yet.

A general conclusion can be that the transition from the 'paper' to the 'digital regime', with traditional publication formats disappearing and new, digital formats replacing these, not only saves on costs for printing and archiving, but also opens possibilities to analyse documents automatically and to increase insight in the way they are used. This can only be beneficial for the legal profession, for which document processing has always been of primary importance.

## References

Albarrán & Ruiz-Castillo 2011

P. Albarrán and J. Ruiz-Castillo, 'References made and citations received by scientific articles', in: *Journal of the American Society for Information Science and Technology*, 62(1), p. 40–49.

ALWD & Darby Dickinson 2010

Association of Legal Writing Directors & Darby Dickerson, *ALWD Citation Manual: A Professional System of Citation*, 4th ed., New York: Aspen 2010).

Campbell 2008

Philip Campbell, 'Escape from the impact factor', in: *Ethics in Science and Environmental Politics* Volume 8 (2008), p. 5-7.

Garfield 1972

Eugene Garfield, 'Citation Analysis as a Tool in Journal Evaluation', in: *Science* 178 (1972), p. 471-479.

Garfield 1983

Eugene Garfield, *Citation Indexing – Its theory and application in Science, Technology and Humanities*, Philadelphia (USA): ISI Press 1983, originally published by John Wiley & Sons, 1979.

Heilbron 2005

Johan Heilbron, *Scientific Research: Dilemmas and Temptations*, Amsterdam: Royal Netherlands Academy of Arts and Sciences, 2005, available online: <https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/knawdilemmasandtemptations.pdf>, consulted 19 November 2014.

Kleinberg 1999

Jon M. Kleinberg, 'Authoritative sources in a hyperlinked environment', in: *Journal of the ACM (JACM)* 46.5, 1999, p.604-632.

Leydesdorff et al. 2011

Loet Leydesdorff, Lutz Bornmann, Rüdiger Mutz and Tobias Opthof, 'Turning the tables on citation analysis one more time: Principles for comparing sets of documents', in: *Journal of the American Society for Information Science and Technology*, Volume 62/7, July 2011, p. 1370-1381.

Moed 2005

Henk F. Moed, *Citation Analysis in Research Evaluation*, New York: Springer 2005.

Narin 1976

Francis Narin, *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*, Washington, D.C: Computer Horizons 1976.

Pritchard 1981

Alan Pritchard, *Bibliometrics: A bibliography and index*, Watford, Herts.: ALLM Books 1981.



#### Seglen 1997

Per O Seglen, 'Why the impact factor of journals should not be used for evaluating research', in: *BMJ* Volume 314, 15 February 1997, p. 498-502.

#### Shapiro 1992

Fred R. Shapiro, 'Origins of Bibliometrics, Citation Indexing, and Citation Analysis: The Neglected Legal Literature', in: *Journal of the American Society of Information Science*, 45(5), p. 337-339, Wiley-Blackwell 1992.

#### Van Opijnen 2014

M. van Opijnen, *Op en in het Web* (On and in the Web, How the Accessibility of Judicial Decisions Can Be Improved), Dissertation University of Amsterdam, Den Haag: Boom Juridisch 2014, 656 pages, with a summary in English.

#### Schuijt 2010

G.A.I. Schuijt (Ed.), *Leidraad voor Juridische Auteurs* (Instructions for legal authors), Deventer: Kluwer 2010.

#### Walter *et al.* 2003

Garry Walter, Sidney Bloch, Glenn Hunt and Karen Fisher, 'Counting on citations: a flawed way to measure quality', in: *Medical Journal of Australia*, 178(6), p. 280-281.

---

[1] See for example Heilbron 2005.

[2] ALWD & Darby Dickinson 2010, p. 3.

[3] Pritchard 1981; Shapiro 1992, p. 337.

[4] Web of Science is currently owned and operated by Clarivate Analytics, previously the IP & Science business of Thompson Reuters.

[5] Garfield 1972, p. 527.

[6] See Narin 1976.

[7] See for example Leydesdorff *et al.* 2011 and Albarrán & Ruiz-Castillo 2011.

[8] For instance Seglen 1997, p. 498; Campbel 2008, p.5; Leydesdorff *et al.* 2011, p. 1371.

[9] Moed 2005, p. 26.

[10] High impact is definitely not equivalent to high quality, as Walter *et al.* (2003) rightly point out.

[11] A notable exception is the journal 'Tijdschrift voor Rechtsgeschiedenis' ('Legal History Review'), ISSN 0040-7585, which is represented in both WoS and Scopus.

[12] See for instance Garfield 1983, p. 12-15.

[13] To illustrate this, of the 201 different legal journals available via the Rechtsorde.nl system that is described in section 3.3, 197 are in Dutch and only 4 (or 2% of the total) in English.

[14] The legal content integration system Rechtsorde.nl, which discloses all Dutch legal sources, contains 32.289 articles (separate documents) of information type 'journal article' and published in 2017. This excludes case law published in periodicals. For more information on content integration, see paragraph 3.3.

[15] Two content integration systems have been used in the preparation of this paper. The first, Legal Intelligence (<http://www.legalintelligence.com>) was introduced in 2004, the second, Rechtsorde (<http://www.rechtsorde.nl>) in 2006. Legal Intelligence is currently a subsidiary of Wolters Kluwer Publishers while Rechtsorde is owned by Sdu Publishers (part of the French company ELS Publishers).

[16] Van Opijnen describes, as an example, the following regular expression that covers all valid references to the publication 'VakstudieNieuws' (by Kluwer), abbreviated as 'VN' or 'V-N', after which either a year and a page number are specified, or alternatively a year, edition number, optional group number and serial number: `(V(-)?N(,|:)?\s\d{2,4}(/|,|s)\d{1,2}(\.D{1,2})*)`

[17] Such recommendations, for Dutch legal publications, can for instance be found in Schuijt 2010.

[18] In fact, in the current version of the Rechtsorde system, the list of related documents has been split into three separate lists: 'Related comments', 'Related case law' and 'Other related publications'. Most of our links appeared in the latter (sub)list.