

# A free access, automated law citator with international scope: the LawCite project

---

Andrew Mowbray, Philip Chung and Graham Greenleaf\*

4 November 2016; 9,936 words

## *Contents*

|  |           |
|--|-----------|
| <b>1. Introduction – Aims of the LawCite project</b> .....                   | <b>2</b>  |
| 1.1 The free access context.....   | 3         |
| <b>2. Legal citator and automated hypertext research in context</b> .....    | <b>3</b>  |
| <b>3. Overall architecture of the LawCite project</b> .....                  | <b>5</b>  |
| <b>4. The data mining processes used by the LawCite project</b> .....        | <b>6</b>  |
| 4.1 Purposes of the data mining processes .....                              | 6         |
| 4.2 Data selection.....  | 6         |
| 4.3 Citation list gathering and parsing .....                                | 6         |
| 4.4 Series database.....   | 8         |
| <b>5. The citation consolidation processes ('unmining')</b> .....            | <b>10</b> |
| 5.1 Purpose of unmining processes .....                                      | 10        |
| 5.2 Citation consolidation .....   | 10        |
| 5.3 Consolidated Output Example .....  | 12        |
| <b>6. The resulting LawCite databases</b> .....                              | <b>14</b> |
| 6.1 Document database.....   | 14        |
| 6.2 Citations database.....  | 15        |
| <b>7. Applications of LawCite's software and databases</b> .....             | <b>15</b> |
| 7.1 Current applications of LawCite tools.....                               | 15        |
| 7.2 The LawCite Citator.....   | 16        |
| 7.3 LawCite as a markup tool .....   | 18        |
| 7.4 The LawCite Markup Tool – a public access markup interface .....         | 20        |
| 7.5 Alternative ranking method for search results By Citation Frequency..... | 21        |
| 7.6 Augmenting Court databases with citation data .....                      | 21        |
| 7.7 Future applications – Research and products .....                        | 21        |
| <b>8. Further development of LawCite tools</b> .....                         | <b>24</b> |
| <b>References</b> .....  | <b>24</b> |
| <b>Appendix 1 – Series database summary</b> .....                            | <b>26</b> |

---

\* Andrew Mowbray is Professor of Law & Information Technology, University of Technology, Sydney, and Co-Director, AustLII; Philip Chung is Associate Professor in Law, UNSW Australia, and Executive Director, AustLII; Graham Greenleaf is Professor of Law & Information Systems, UNSW Australia and Founding Co-Director and Senior Researcher, AustLII. Thanks to Nicholas Tobias for his theoretical work on citation identification. We have received very helpful comments on article drafts from Marc van Opijnen, Judge Grant Reithmuller, two anonymous referees, and the Editors. All responsibility for content, however, remains with the authors. Research for this article was supported by Australian Research Council Linkage Grant LP130100382.

## **1. Introduction – Aims of the LawCite project**

The LawCite Project developed by the Australasian Legal Information Institute (AustLII),<sup>1</sup> since 2008 (Mowbray, Chung and Greenleaf, 2009), aims to maximise the value of documents located on, or known about by, free access legal information institute (LIIs) involved in the project.

The principal application of the project to date is the LawCite citator,<sup>2</sup> which currently contains index records of the citation histories of almost five million<sup>3</sup> cases, law journal articles, law reform documents and treaties. The citator is international, containing citation records in significant numbers from court decisions in 75 countries (primarily but not exclusively from common law countries). The citator is free access to all users, and built for and by thirteen non-profit legal information institutes (AustLII and other collaborating LIIs),<sup>4</sup> which also have access to the project's other resources. This free-access context provides benefits – and imposes constraints – which make it unique.

The LawCite citator and the databases from which it is generated have been built by entirely automated means without editorial intervention, using data mining techniques based on heuristic recognition of citations in source documents. Although the citator is the first and most visible product of the project, the data mining techniques used, and the data sets generated by their use, have other valuable applications.

The project's aims, to put them in slightly more detail, are to maximise the value both of documents located on the collaborating LIIs, and of documents which they do not hold but to which their other documents refer. This is done by recognising and exploiting interconnections between citation information about such documents. The resulting applications improve the functionality of all participating LIIs, particularly in relation to comparative research.

The purpose of this paper is to explain how the components of the LawCite project work, in some technical detail, and to outline the applications that have already resulted (both the citator and others), and future applications that are planned and possible. Our goal is for the LawCite project to play a key role in future global development of free access to legal information, provided collaboratively by free access providers across the globe.

---

<sup>1</sup> AustLII, a free access non-profit facility operated by two Law Faculties, is Australia's largest online provider of legal information. It is consistently rated by the Hitwise Internet ranking service as Australia's most used online legal service, receiving over 30% of all internet traffic in Australia related to legal research. AustLII receives on average 700,000 page accesses per day, measured from its logs. AustLII is a non-profit free-access joint facility of the Faculties of Law at the University of New South Wales and the University of Technology, Sydney..

<sup>2</sup> LawCite citator user interface <<http://www.austlii.edu.au/lawcite/>>.

<sup>3</sup> 4,822,408 records as at 2 April 2016.

<sup>4</sup> In alphabetic order: AsianLII, AustLII, BAILII, CanLII, CommonLII, Cylaw, HKLII, LiberLII, LII of India, NZLII, PaCLII, SAFLII and WorldLII.

### 1.1 The free access context

The fact that all the participants in the LawCite project are free access and non-profit providers of legal information, and that end-user access to the project applications is free of any user charges, imposes constraints on project development. First, only modest funding<sup>5</sup> was available to fund the initial project infrastructure, and the first application (the citator). Open source software plus purpose-developed software was therefore used. Second, no significant editorial input has been available to develop application data, everything has had to be automated. Third, the participating LIIs can only provide very limited maintenance funds (mainly from AustLII). Fourth, no user charges or advertising revenue have been available to support any of these costs. Finally, the participating LIIs do not consider they are competing with each other: the LawCite project is a collaborative effort for mutual benefit. At present, 13 LIIs provide their data for use in the project.

## 2. Legal citator and automated hypertext research in context

A number of free access law providers have produced or are developing citators. The RefLex citator developed by Lexum and used on CanLII focuses on case law and primarily Canadian content (Poulin, Pare and Mokonov, 2005). Indian Kanoon focuses on Indian case law.<sup>6</sup> Caselaw NSW has editorially tracked cases which cite law reform reports.<sup>7</sup> Some Legal Information Institutes (eg AustLII) use mark-up software to provide hypertext linking to current legislative sections and to cases which cite other cases by a court-designated citation, but not by the parallel 'publishers citations' still most commonly used. Other free access sources such as Google Scholar and SSRN/LSN focus on law journal articles and commentaries (which LawCite also includes), rather than on case law. None of these projects have both the international focus of the LawCite project, and none cover both case law and legal scholarship, as does LawCite.

Commercial legal publishers often achieve functionality similar to LawCite by largely editorial means exerted over many years at great expense. They may also use proprietary automated means that are not public. This editorial value-adding is a significant part of the services that commercial legal publishing provides. Examples are 'Shepardizing' or 'Keyciting' cases (equivalent of what we call 'noting up') and 'case citators' which identify parallel citations, appeal status of decisions, and editorial comments on how cases are interpreted by later cases. This project seeks to achieve as much of that value-adding as possible by low cost largely automated means, and to share the results.

There is limited literature on heuristic-based automated linking of legal materials (Rugh and Lennen, 2003 uses fuzzy logic instead). Outside law, there is

---

<sup>5</sup> LawCite development, including both infrastructure development and expansion of content indexed via LawCite, has been one component of a number of Australian Research Council (ARC) grants obtained since 2008, including two Linkage grants and a number of Linkage Infrastructure, Equipment and Facilities (LIEF) grants.

<sup>6</sup> Indian Kanoon <<https://indiankanoon.org/>>.

<sup>7</sup> Caselaw NSW <<https://www.caselaw.nsw.gov.au/policy.html>>.

a knowledge base on the use of heuristics and other techniques for citation recognition. While it is useful it does not solve the problems posed by legal information or even address many of them. For scientific and technical literature, an approach known as autonomous citation indexing (ACI) has been effectively implemented in the CiteSeer system<sup>8</sup> (Lawrence, Giles, and Bollacker, 1999; Giles, Bollacker, and Lawrence, 1998). While the CiteSeer conceptual framework might prove useful in the legal domain, its algorithms are grounded in scientific and technical documents. As detailed herein, many more variations of case citations are possible, and the information needed for CiteSeer to work effectively may not be available or be as consistent and obvious in legal documents. For example, cases are often cited with page numbers missing, without a date, by different names, with different formatting and punctuation, with typographical errors, or with nothing but the case name. Thus “Mabo and Others v The State of Queensland (1992) 175 CLR 1” may be cited as “Mabo & Ors -v- Queensland 175 CLR, 1”, “Mabo v Queensland (1992) 175 C.L.R.”, or simply “Mabo”.

The problem of resolving legal citations cannot be solved by simple application of existing algorithms. For example, the Naïve Bayes Classifier<sup>9</sup> approach fails in the legal domain because of the complex web of interdependencies between the various elements of legal citations. Whereas more correlations between components of non-legal citations generally indicate a greater likelihood of a match, this does not apply for legal citations. For example, “175 CLR 1” refers to the same case as “[1992] HCA 23”, but not the same case as “174 CLR 1”. Further, two cases may have the same name and occur in the same year and jurisdiction and yet be distinct from one another. This makes a Naïve Bayes approach unsuitable. Borkowski (1969) developed an algorithm for extracting legal citations. However, it was only tested on formal citations with no variations in structure or format. The algorithm produced extremely poor results when tested with cases in AustLII’s database, due to the many variations in format, structure and style of today’s legal citations. Consequently, his algorithm is only of limited use.

Fuzzy logic approaches have been used to resolve legal citations in the past, but these have used complete, high quality, manually entered data sets. Because AustLII does not have the resources to enter all parallel citations manually, it must obtain data by isolating and extracting legal citations from cases through automated means. Consequently, the data is of much lower quality and so the fuzzy logic does not work well. Other approaches for citation resolution combine extraction and co-reference so as to obtain a higher quality of data (Rugh and Lennen, 2003), and use overlapping canopies to efficiently cluster citations (McCallum and Ungar, 2000). These approaches do not transpose all that well into the legal domain.

---

<sup>8</sup> CiteSeer system at <<http://citeseer.ist.psu.edu/directory.html>>.

<sup>9</sup> For an introduction, see ‘Naive Bayes classifier’ (Wikipedia entry) <[https://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)>: ‘In [machine learning](#), naive Bayes classifiers are a family of simple [probabilistic classifiers](#) based on applying [Bayes’ theorem](#) with strong (naive) [independence](#) assumptions between the features.’

While elements of some of these approaches have been used in AustLII's algorithms and heuristics used to develop LawCite, our approach is law specific.

### 3. Overall architecture of the LawCite project

The LawCite project has developed and uses 3 main databases in an iterative fashion to produce its applications (the citator and other applications). They are:

- The Citations database (basic information related to a citation )
- The Series database (information about each series of law reports, law journals, treaties, or law reform reports recognised by LawCite)
- The Document database (an XML record for each recorded case or journal article)

Oracle Berkeley DB (database software) is used for data mining and markup, and for construction of the Citations and Series databases. The Series and Document databases are used to generate the Citations database. The Series database is manually updated periodically. The Document and Citations databases are recreated automatically during each iteration of the data mining and 'unmining' processes next described.

The following diagram shows the overall system design and the main components of the LawCite project.

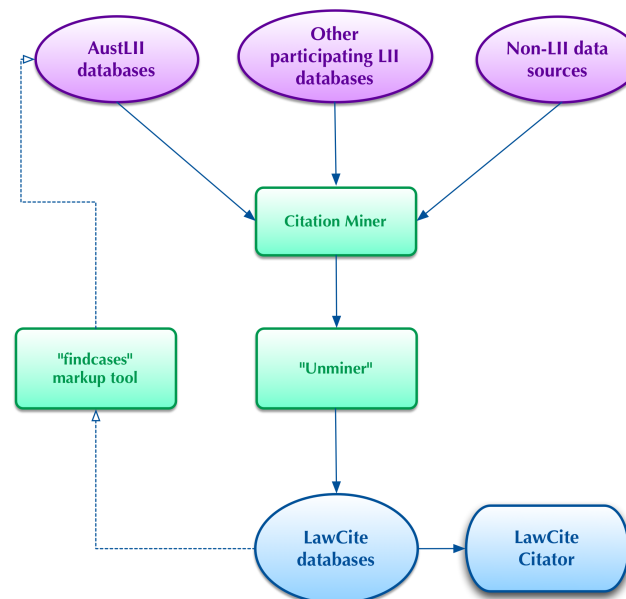


Figure 1 Main components of the LawCite project

Raw citation lists are gathered from participating LII and non-LII data sources by the Citation Miner. Once collected, these are analysed, combined and normalised by the "Unminer" and from the combined list, the LawCite databases are generated. These databases are used both by the LawCite Citator as well as for the markup of text.

## **4. The data mining processes used by the LawCite project**

### **4.1 Purposes of the data mining processes**

Major tasks in the construction and generation of the Citations database and the Document database include:

- Data selection
- Citation list gathering / parsing
- Citation consolidation
- Database creation

The following sections give a more technical explanation, describing the algorithms and heuristics used in each phase of the Citations database generation process.

### **4.2 Data selection**

The first phase is selecting which data to mine. At the moment, the primary resource is the WorldLII database which includes most of the content published by the members of the Free Access to Law Movement. Other internet content, for example, the Supreme Court of Canada and other major case law databases that are not included on WorldLII are also mined for citations.

The files include texts of cases, journal articles, law reform reports and treaties, and any other documents that contain citation lists. Currently, over 4 million files are mined for citation information.

The list of files to mine is generated by the LawCite Citation Miner. This process avoids duplication and excludes data known to be unreliable through the use of an exclusions list. Once the files have been identified, each of these is analysed via a number of concurrent processes.

### **4.3 Citation list gathering and parsing**

For each mined document, this phase identifies every individual reference to a case, journal article or law reform report, both by name and citations to other cases, journal articles and law reform reports. Where a mined document is a case, journal article, law reform report or treaty available on an LII, appropriate meta-data is also included in record citation lists. With the current dataset, about 34 million raw citation lists are generated.

A citation list includes a reference to a case or article by name followed by a set of parallel citations. For LII files, the date, jurisdiction and file location are also recorded. Heuristics are used to attempt to parse valid case names. A case name must have a valid start term and then contain a number of capitalised words. There are various English words and some other exceptions that are allowed (eg use of words like “and”, “the”, “of” and so forth). A case name must generally also take a valid overall form. These include formats like X v Y, Re X and so on.

A parallel citation list consists of a number of citations separated by a valid connector (usually a semicolon). The citation itself must be in a valid overall form. Typically, a citation starts with a year (that may be optional and possibly in round or square brackets), then an optional volume, a series and a page

reference or decision number. There are a number of extensions to this format to allow for various non-standard citations. The parser for these citations is rule based and can be changed to allow for new citation formats.

The following shows what LawCite considers to be a 'citation' using an EBNF-like<sup>10</sup> notation.

```

“LawCite citation structure” {
    citation      = ( year | [ year ] volume_issue ) series page [ division ] .
    year          = number | ‘(‘ number ‘)’ | ‘[‘ number ‘]’ .
    volume_issue = ‘(‘ volume ‘)’ | volume ‘(‘ issue ‘)’ .
    volume       = number .
    issue        = number .
    series       = capword { capword | connectword } . (* max 30 words *)
    page         = digit { digit | ‘-’ | ‘,’ } .
    division     = ‘(‘ capital { character } ‘)’ .
    capword      = capital { character } | ‘e’ word .
    connectword = ‘in’ | ‘the’ | ‘for’ | ‘on’ | ‘and’ | ‘of’ .
    capital      = ‘A’ | ‘B’ | ‘C’ | ‘D’ | ‘E’ | ‘F’ | ‘G’ | ‘H’ | ‘I’ | ‘J’
                | ‘K’ | ‘L’ | ‘M’ | ‘N’ | ‘O’ | ‘P’ | ‘Q’ | ‘R’ | ‘S’ | ‘T’
                | ‘U’ | ‘V’ | ‘W’ | ‘X’ | ‘Y’ | ‘Z’ .
    digit        = ‘0’ | ‘1’ | ‘2’ | ‘3’ | ‘4’ | ‘5’ | ‘6’ | ‘7’ | ‘8’ | ‘9’ .
    number       = digit { digit } .
    word         = character { character } .
} “get_case_ref() -- parse an individual case/journal citation.”

```

The parser is currently best suited to find common law case citations. For discussion of how this could be extended to deal with civil law jurisdictions, see for example, van Opijnen, Verwer and Meijer (2015).

The following example shows part of the metadata (including case and article references) extracted from the well-known Australian case of *Mabo v Queensland*:

```

# file /home/www/au/cases/cth/HCA/1992/23.html
|19920603|AU|/home/www/au/cases/cth/HCA/1992/23.html|Mabo v Queensland (No 2) ("Mabo
case")|[1992] HCA 23|(1992) 175 CLR 1
||AU||Wade v. New South Wales Rutile Mining Co. Pty. Ltd|(1969) 121 CLR 177
||AU|||[1959] HCA 63|(1959) 102 CLR 54
||AU|||[1975] 135 CLR 337
||AU|||[1915] 20 CLR 425
||AU|||[1892] AC 437
||AU||Jones v. The Commonwealth|(1987) 61 ALJR 348
||AU|||71 ALR 497
||AU||John v. Federal Commissioner of Taxation|(1989) 166 CLR 417
||AU||McKinney v. The Queen|(1991) 171 CLR 468
||AU||Sobhuza II. v. Miller|(1926) AC 518

```

These are raw lists of citations that may contain incomplete references, conflicts and other errors.

<sup>10</sup> 'In computer science, extended Backus–Naur form (EBNF) is a family of metasyntax notations, any of which can be used to express a context-free grammar.: 'Extended Backus–Naur form' (Wikipedia entry) <[https://en.wikipedia.org/wiki/Extended\\_Backus%E2%80%93Naur\\_form](https://en.wikipedia.org/wiki/Extended_Backus%E2%80%93Naur_form)>.

## 4.4 Series database

### 4.4.1 Purposes

In addition to being in appropriate form, a citation must use a legal series identifier (eg a law report series, a court database, a law journal). This database provides information about preferred series abbreviations, whether or not a year is required and so forth. The name of a series is normalised and compared against a large series database.

The series database contains over 18,000 valid series and series abbreviations of media neutral and conventional law reporting series, law journals, law reform reports and treaties. The main source used to construct the series database is the Cardiff Index to Legal Abbreviations<sup>11</sup>, but other citation lists are also used.<sup>12</sup> The current emphasis is on common law countries, but this is being gradually extended to include civil law jurisdictions.

### 4.4.2 Format

Each entry is either a substantive entry or an abbreviation of a substantive entry. A substantive entry contains information about the series type eg journal, case etc., the full name of the series, the court or tribunal, the country, the sub-jurisdiction (if applicable), the publisher and the destination or location (if available).

For example, the following is the entry for the neutral citation of the High Court of England and Wales:

```
EWHC|EWHC|cases division-required neutral|High Court of England and Wales|BAILII|High
Court of England and Wales|UK|England and Wales
|http://www.bailii.org/ew/cases/EWHC/$(division)/$(year)/$(decno).html
```

The first field is the compressed or normalised name of the series. The second field is the output format for the series. The third field contains the series type (case) which requires special parsing (ie that includes a court division) and that it is a neutral citation. The fourth field contains the full name of the series ('High Court of England and Wales'). The decisions of this series are available on 'BAILII'. The series contains only decisions from the High Court of England and Wales. The next two fields contain the country code ('UK') and where applicable, the sub-jurisdiction ('England and Wales'). The final field contains the meta rule for determining the actual location of decisions in a particular vendor neutral series. In this instance, the decisions are on BAILII in the 'EWHC' directory followed by the 'division' and 'year' and the filename is based upon the decision number ('decno').

---

<sup>11</sup> Cardiff Index to Legal Abbreviations < <http://www.legalabbrevs.cardiff.ac.uk/> > "This database allows you to search for the meaning of abbreviations for English language legal publications, from the British Isles, the Commonwealth and the United States, including those covering international and comparative law. A wide selection of major foreign language law publications is also included. Publications from over 295 jurisdictions are featured in the Index. The database mainly covers law reports and law periodicals but some other legal publications are also included."

<sup>12</sup> These include citations lists from [Monash Law Library](#), the Melbourne University MULR [Australian Guide to Legal Citation](#) and the [University of New South Wales Law Reports and Abbreviations Database](#).



The entries for commercially published series tend to be simpler. For example:

```
F3D|F3d|cases respected|Federal Reporter, Third Series|Westlaw||US||
```

Entries for law journal series have the same fields. The following is the entry in the series database for the African Human Rights Law Journal.

```
AFRICANHUMANRIGHTSLAWJOURNAL|African Human Rights Law Journal|journals issue-
required|African Human Rights Law Journal|HeinOnline / Juta|African Human Rights Law
Journal|ZA||
```

#### 4.4.3 Fields and flags

The fields and flags used in the above content formats are:

- The type field: 'cases', 'journals', 'lawreform' – default: 'case'
- Parsing flags: 'european', 'complex-page', 'division-required', 'issue-required', 'year-required'
- The rank order flag: 'respected', 'authorised', 'neutral' – default: conventional law report series
- The presentation order flag: 'neutral', 'authorised', 'respected'

The information contained in the series database will continue to expand. Some of the additional elements that are anticipated include: valid time intervals (eg a commercial series came into existence and then ceased publication), more differentiating parsing requirements (eg citations within a particular jurisdiction), and more flexible overall citation templating.

#### 4.4.4 Resulting content

Of the over 18,000 valid series and series abbreviations in the series database, the top five countries based on the total number of series recognised by LawCite are listed in the table below. A complete list of all countries (158) is provided in Appendix 1. The entries counted in the case law series and journal series fields include both the neutral citation and the printed citation of a particular publication.

**Table 1 Top five countries in the LawCite Series database (by number of series)**

| Country        | Case Law Series | Journal Series | Total Number of Series |
|----------------|-----------------|----------------|------------------------|
| United States  | 1967            | 1888           | 3855                   |
| United Kingdom | 1330            | 425            | 1755                   |
| Australia      | 665             | 458            | 1123                   |
| Canada         | 786             | 186            | 972                    |
| India          | 417             | 98             | 515                    |
|                | <b>5165</b>     | <b>3055</b>    | <b>8220</b>            |

The top five countries' combined total number of series constitutes the bulk of the entries in the series database which is about 78% of the total number of

series in the database. These account for about 75% of the case law series entries and 84% of the journal series entries in the database.

## 5. The citation consolidation processes ('unmining')

### 5.1 Purpose of unmining processes

From the raw lists of citations extracted in the previous phase, which may contain conflicts and other errors, the LawCite Citation 'Unminer' attempts to resolve conflicts and other errors, and to consolidate lists of citations.

Once identified, the data miner outputs a raw set of citation lists for each file that has been mined (as well as outputting and identified meta-data for the file itself). The number of raw citation lists produced by the data mining process is large (around 32 million at the moment). Whilst each list has already had some checking, an individual case or journal article will generally be represented by multiple entries, some of which can be combined, whilst other will lead to 'conflicts'. A conflict is any situation where two citations cannot coexist.

A major problem that the system has to deal with is the presence of incorrect citations and citation sets in the source data (mainly judgments, but also journal articles and law reform reports). Conflict identification and correction mechanisms have been developed to deal with this.

Consider the following example of a set of references that might be identified for the case *Dietrich v The Queen*:

```
Dietrich v The Queen [1992] HCA 57; (1992) 177 CLR 292
Dietrich v The Queen [1992] HCA 57; (1992) 177 CLR 291
Dietrich v The Crown [1992] HCA 57; (1992) 177 CLR 292
Dietrich v R [1992] HCA 57; (1992) 177 CLR 292
Dietrich v The Queen [1991] HCA 57; (1992) 177 CLR 292
(1992) 177 CLR 292; [1992] HCA 77
Dietrich v R [1992] HCA 57; (1992) 177 CLR 292; [2000] 2 All ER 12
```

This set contains a number of typographical and other errors (indicated in bold) including incorrect court designated citations, an error in the publishers' citation and a stray reference to an unrelated English decision. Note also that the name of the case varies. The LawCite conflict resolver deals with all of these entries by (i) counting the most frequent associations for each individual citation, (ii) rejecting conflicts based upon reports from different jurisdictions or courts and (iii) tracking the most commonly used case name. The above example set is reduced to:

```
Dietrich v The Queen [1992] HCA 57; (1992) 177 CLR 292
```

While the approach is highly heuristic, we believe it achieves high levels of accuracy that are rapidly becoming comparable or better than manual editorial approaches. Further testing will be needed to demonstrate this.

### 5.2 Citation consolidation

The simplified consolidation algorithm used by the LawCite Citation Unminer is:

```
For each citation set / file
```

```

    Record legal set against each citation
  For each individual citation
    Resolve conflicts for matching citation list
  For each unprocessed citation
    Fully resolve and mark as processed and add to output set
  For each entry on the output set
    Check if this is a typo
  Output the output set

```

The ‘unmining’ process involves seven stages, each described below.

### **5.2.1 Stage One: read in the case name / location / citation lists**

An in-memory table is built for each individual citation found in any of the input lists. Each possible case or journal article name is recorded and counted, as follows:

- ignore comments
- ignore lines that do not parse as a valid entry
- ignore lines that contain conflicts if in conservative mode
- record the entry in the citation table

### **5.2.2 Stage Two: clean up each citation list to make consistent**

Each of these lists is resolved so as to be internally consistent, and so that it uses the most commonly occurring consistent citations. There are a number of conflict situations that must be resolved, including:

- conflicting types – an entry cannot be both a case and a journal article;
- conflicting years – years must be consistent;
- conflicting jurisdictions – jurisdictions, both at a country and sub-country level, must be consistent;
- conflicting courts – courts, where determinable from the citation, must be the same;
- conflicting with a series – a case can only appear at one point in a series.

Where a conflict occurs, this is generally resolved by eliminating the minimum number of least frequently occurring entries to render the list consistent, by the following steps:

- resolve conflicts in the citation list
- mark the most popular citation format for matches
- mark conflicts – citations may ‘conflict’ in the following ways: case/journal conflict, that is, type conflict; year conflict; country conflict; jurisdiction conflict; court conflict; and series conflict
- remove bad / unnecessary entries

### **5.2.3 Stage Three: check for ambiguous citations**

Once a complete set of internally consistent citation lists has been built, ambiguous citations are identified and marked. A citation is ambiguous where it appears on two lists that cannot be combined.

#### 5.2.4 Stage Four: fully resolve citations and output

The citation lists are merged and fully resolved. Again, conflict checking is done to produce the most likely set of unique lists. Ambiguous citations are allowed to remain on multiple lists.

#### 5.2.5 Stage Five: remove / redirect typos

The system attempts to eliminate typos. A 'typo' is a case that occurs relatively infrequently that shares a name with a much more commonly used case, and that can be mapped to a similar citation. A 'similar' citation is one that differs by any single element (a year, volume or page) only. No output record is generated for such instances.

#### 5.2.6 Stage Six: output the results

The resultant lists are output to (re-)build the Citations database. This is the used in applications for markup and to build the Document database of XML files for the LawCite Citator.

#### 5.2.7 Stage Seven: tidy up, report statistics and exit

Finally, statistics are generated (see diagram below, S7).

### 5.3 Consolidated Output Example

After applying the consolidation algorithm described above to the raw citation lists generated in the mining process, consolidated lists of citations are created. These form the basis for the construction of the Citations database. An example follows:

```
48|19900416|US|/us/cases/federal/USCA7/1990/296.html|Yeksigian v. Nappi|[1990] USCA7
296|900 F2d 101
27|19901121|US|/us/cases/federal/USCA7/1990/1036.html|Town of Concord Massachusetts v
Boston Edison Co|[1990] USCA1 426|915 F2d 17|59 USLW 2201
3|19900614|US|/home/raid2/data/us/cases/federal/USCA7/1990/502.html|Anonymous v
Anonymous|907 F2d 152
12|19900920|US|/home/raid2/data/us/cases/federal/USCA7/1990/851.html|Cohen v
Commissioner of Internal Revenue|[1990] USCA7 851|910 F2d 422|66 AFTR2d 90-6004
```

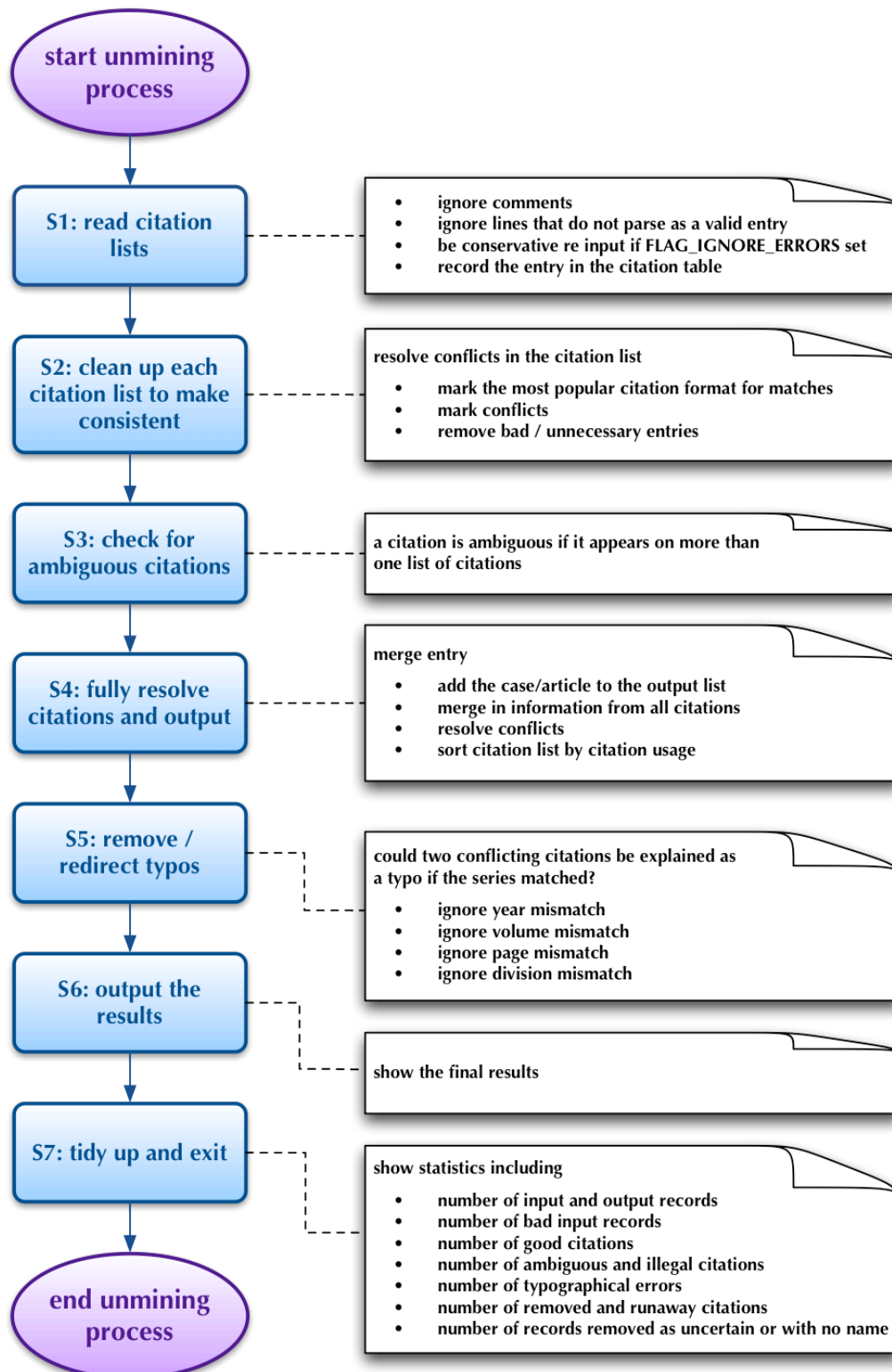


Figure 2 Seven stages of the LawCite 'unmining' process

## 6. The resulting LawCite databases

### 6.1 Document database

The Document database contains an XML record for every document (case or journal article) referred to in any document that is mined by LawCite. There are currently close to five million records.<sup>13</sup> The following is an example of a typical XML record.

```
<case>
  <title>
    Kable v DPP (NSW) [1996] HCA 24; (1996) 189 CLR 51; 138 ALR 577; (1996) 70 ALJR
    814; (1996) 14 Leg Rep C1; [1996] 3 CHRLD 435 (12 September 1996)
  </title>
  <name> Kable v DPP (NSW) </name>
  <court> High Court of Australia </court>
  <jurisdiction> Australia - Commonwealth </jurisdiction>
  <date> 12 September 1996 </date>
  <url> http://www.austlii.edu.au/au/cases/cth/HCA/1996/24.html </url>
  < Citations >
    < citation > [1996] HCA 24 </ citation >
    < citation > (1996) 189 CLR 51 </ citation >
    < citation > 138 ALR 577 </ citation >
    < citation > (1996) 70 ALJR 814 </ citation >
    < citation > (1996) 14 Leg Rep C1 </ citation >
    < citation > [1996] 3 CHRLD 435 </ citation >
  </ Citations >
  < legislation >
    < section >
      < number > 5 </ number >
      < url > http://www.austlii.edu.au/au/legis/act/consol_act/ca190082/s5.html
      </ url >
    </ section >
  < act >
    < short-title > Australia Act 1986 (Cth) </ short-title >
    < url >
      http://www.austlii.edu.au/au/legis/cth/consol_act/aa1986114/index.html
    </ url >
  </ act >
  ...
</ legislation >
< cases >
  < cited >
    Australian Consolidated Press Ltd v Uren [1969] 1 AC 590
  </ cited >
  < citedref > 1969AC590 </ citedref >
  ...
```

The key element of each of these records is a list of cases, journal articles and legislation that it cites (provided that the text of the document is available). These records include other metadata about each case or journal article including the title, jurisdiction, date and the web address of the document (if available).

There are over 4.3 million case records and over 450,000 journals and law reform records in the LawCite Document database.

<sup>13</sup> The current number is shown on the LawCite citator page < <http://www.austlii.edu.au/lawcite/>>. As at 18 February 2016, it is 4,809,684 indexed cases, law reform documents and journal articles.

## 6.2 Citations database

The Citations database provides a different view of the documents included in LawCite. Its key role is to provide a fast mechanism to support the insertion of hypertext links by the associated markup software.

Examples:

```
35LED2D147|3159|19730226|US|http://www.worldlii.org/us/cases/federal/USSC/1973/43.html|Roe v.
Wade|[1973] USSC 43|410 US 113|93 Sct 705|35 LEd2d 147
146PHIL469|3||PH||Edu v. Ericta|146 Phil 469
208PHIL151|6||PH||Soco v. Militante|208 Phil 151
131PHIL1022|2||PH||Caltex Filipino Managers & Supervisors Ass'n v. CIR|131 Phil 1022
```

The first field contains the compressed or normalised citation. The second field contains the absolute citation count of the document. The third field contains the date (if available). The fourth field contains the country code. The rest of the fields contains the location, the title of the document, and parallel citations.

## 7. Applications of LawCite's software and databases

### 7.1 Current applications of LawCite tools

The current applications of the LawCite project's tools are: (i) the public access LawCite citator; (ii) use of the XML Citation database by AustLII and by other LIIs, via an API, to improve the mark-up of their databases (insertion of additional citations and/or hypertext links); (iii) a public access interface, the LawCite Markup Tool, enabling anyone to mark up their own documents or text by insertion of additional citations and/or hypertext links; and (iv) an additional means of ranking documents retrieved in searches, By Citation Frequency, used by AustLII and other LIIs.

The relationships between LawCite tools and applications is as follows:

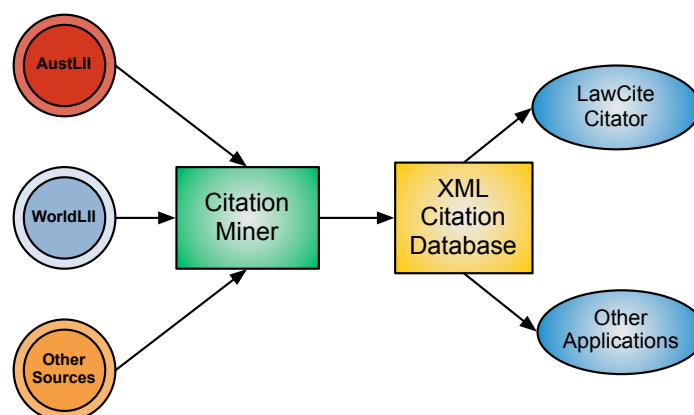


Figure 3 LawCite tools and applications

## 7.2 The LawCite Citator

The LawCite citator is the public access interface<sup>14</sup> to the Citations database. It provides information on nearly 5 million indexed cases, law reform documents and journal articles.

LawCite’s search screen (below) allows numerous ways to find cases or articles. To find cases, a citation, or any combination of party name(s), court, jurisdiction and/or year(s)<sup>15</sup> may be used. The name of one party, and the court (or sometimes even the jurisdiction), can also be used.

Figure 4 LawCite Citator homepage

LawCite search results show all cases,<sup>16</sup> law reform reports and law journal articles matching the search, sorted in default by frequency of citation (ie by the Citation Index field). The search results can be sorted by selecting any of the other fields in the search results display (for cases: Case Name; Citation(s); Court; Jurisdiction; Date; Full Text; Citation Index). From the list of search results, the user chooses the result item that best matches the search (for oft-cited items, they will often be ranked first in the display). Citation information for that item is then displayed (the ‘LawCite record’ for that item).

### 7.2.1 LawCite records for cases and articles

LawCite Case and Article records consist of a header followed by up to five tables. The header lists the name of the case or article, the citations list, the court/journal, the jurisdiction and the date. Alternative (parallel) citations for the case are displayed with its title. Where a free full-text version of the case, article or law reform report is available the citation in the citations list will be a live link (blue). The following tables (in order) are available:

<sup>14</sup> < <http://www.austlii.edu.au/LawCite/> >



<sup>15</sup> To limit cases found to those from a particular period, users put two years in the ‘Year’ field.

<sup>16</sup> For large lists of search results (more than 5000) the user is prompted as to whether or not they want to continue prior to the list of results being displayed.



- Cases Referring to this Case or Article
- Law Reform Reports Referring to this Case or Article
- Journal Articles Referring to this Case or Article
- Legislation Cited
- Cases and Articles Cited

A LawCite record is in Figure 5, a UK case from 1884 concerning cannibalism on the high seas and the defences of duress and necessity. It is worth noting that of the most recent 8 of 78 known citations, three are from the UK, the Solomon Islands and the International Tribunal for Sierra Leone. The citations also include six law reform reports (from Ireland, Scotland and Australia) and many law journal articles.

**R v Dudley and Stephens**  **78**  [Help](#)

(1884) 14 QBD 273; (1884) 54 LJM 32; (1884) 52 LT 107; (1884) 49 JP 69; (1884) 33 WR 347; (1884) 15 Cox CC 624; (1884) 1 TLR 118

Queen's Bench  
United Kingdom  
9th December, 1884

**Cases Referring to this Case**

















| Case Name  | Citation(s)  | Court  | Jurisdiction                             | Date †      | Full Text                | Citation Index  |
|--|--|--|--|-------------|--------------------------|---|
| <a href="#">Leichhardt Council v Geitonina Pty Ltd (No 6)</a>  | [2015] NSWLEC 51   | Land and Environment Court of New South Wales        | Australia - New South Wales              | 2 Apr 2015  | <a href="#">AustLII</a>  |   3     |
| <a href="#">Nicklison, R (on the application of) v A Primary Care Trust</a>  | [2013] EWCA Civ 961;<br>[2014] 2 All ER 32                 | England and Wales Court of Appeal - Civil Division   | United Kingdom - England and Wales       | 31 Jul 2013 | <a href="#">BAILOI</a>   |   1     |
| <a href="#">R v ML</a>   | [2013] ACTSC 32  | Supreme Court of the Australian Capital Territory    | Australia - Australian Capital Territory | 1 Mar 2013  | <a href="#">AustLII</a>  |     |
| <a href="#">R v Abdulla</a>  | [2010] SASC 52;<br>(2010) 200 A Crim R 365                 | Supreme Court of South Australia                     | Australia - South Australia              | 11 Mar 2010 | <a href="#">AustLII</a>  |   8 |
| <a href="#">Tran v The Commonwealth</a>  | [2009] FCA 474;<br>(2009) 108 ALD 531                      | Federal Court of Australia                           | Australia - Commonwealth                 | 15 May 2009 | <a href="#">AustLII</a>  |   4 |
| <a href="#">Johnson v Western Australia</a>  | [2009] WASCA 71;<br>40 WAR 116;<br>(2009) 194 A Crim R 470 | Supreme Court of Western Australia - Court of Appeal | Australia - Western Australia            | 2 Apr 2009  | <a href="#">AustLII</a>  |   4 |
| <a href="#">Luavex v R</a>   | [2007] SBCA 13   | Court of Appeal of Solomon Islands                   | Solomon Islands                          | 18 Oct 2007 | <a href="#">PacLII</a>   |   1 |
| <a href="#">Prosecutor v Moinina Fofana Judgement on the Sentencing of Moinina Fofana and Allieu Kondewa Case Noscsi-04-14-t</a> | [2007] SCSL 10   | Special Court for Sierra Leone                       | International - Sierra Leone             | 9 Oct 2007  | <a href="#">WorldLII</a> |     |

Figure 5 Example of a LawCite citation record: *R v Dudley and Stephens*

An individual LawCite record can also be sorted on the basis of the contents of each column by clicking on column name at the top of the table. The currently selected column for sorting is indicated by a † symbol. Hovering the mouse over various table elements provides further information about the item or a note as to what will happen if the link is selected.

The columns are as follows:

- *Case/Article/Legislation Name* – This column contains the case name or article title. This is the way the case or article is most commonly referred to in the full-text data from which LawCite is built. If you click on this name, you will get the LawCite entry for this case or article (if one is available).

- *Citations* – The Citations column contains all known citations for a case or article. Clicking on any live (blue) citations will bring up the full-text of the case or article. You can also hover over a citation to see what it is (ie which journal or series of law reports). The order of citations is neutral citation first (if any), authorised citation next (if any) and then citations ordered by how often they have been referenced.
- *Court/Author* – This column contains the name of the court that handed down the decision or the author of an article. This is determined for cases, on the basis of the series that appear in the citations list and in the case of articles, from meta-data in any linked full-text version. This information will not always be available.
- *Jurisdiction* – The Jurisdiction column list the country and (sometimes) the sub-jurisdiction. It is almost always available and again is derived from the series.
- *Date* – This column lists the date or year of a decision or judgment. Dates refer to the date that a decision was handed down. A year is the year that a decision or journal article was first published.
- *Full Text* – The Full Text column lists a place where the full text of the decision or article can be found (for free access). Where available, this will normally be one of the Free Access to Law Movement sites (AustLII, BAILII, HKLII etc). Where no free version exists, the location of the decision on one of the commercial services may be listed, intended only as a non-comprehensive guide to possible commercial locations.
- *Flags* – This column has a flag indicting the nationality of the decision or article. It is intended to make it easy to see at a glance the national origin of results.
- *Citation Index* – The Citations List column contains (as a minimum) a small LawCite logo. Clicking on this will bring up the LawCite record for the case or article. Where the case or article is frequently cited, a number of “stars” will appear. The greater the number of stars, the more frequently the case or article has been cited.

LawCite records do not explicitly include appeal details, although this can often be inferred from the search results for the parties to a case. They do not include editorially-developed information such as ‘distinguished’ and ‘not followed’, as these are not possible in a fully-automated citator. This is an inevitable consequence of the constraints of a free-access, non-profit, citator. The extent to which these editorial enhancements are valuable is open to debate.

### 7.3 LawCite as a markup tool

LawCite includes a simple API to efficiently add citation links based on the LawCite Citations database. This relies on a fast lookup mechanism provided by the Citations database. When presented with a non-neutral citation,<sup>17</sup> the markup tool will insert a neutral citation where this exists and create a link to the file if it is available on a participating LII. It will also create a link from the non-neutral citation to a LawCite record (if available).

---

<sup>17</sup> Citations other than such neutral citations used by the courts themselves or citations commonly used by commercial publishers.

### 7.3.1 Markup example

If the following text is used as source input to the LawCite mark-up software, the resulting mark-up text shown below appears as output.

renewed vigour has been afforded to this approach by its recent restatement in a series of decisions by the House of Lords and the Privy Council. (See *Lim Chin Aik v R* [1963] AC 160 (PC) at 172; *R v Warner* [1969] 2 AC 256 (HL (E)) at 271-2; *Sweet v Parsley* [1970] AC 132 (HL (E)) at 163; ...

... See *Maher v Musson* (1934) 52 CLR 100 at 104-5; *Thomas v R* (1937) 59 CLR 279 at 287-8; *Proudman v Dayman* (1941) 67 CLR 536 at 540; *Iannella v French* (1967-1968) 119 CLR 84 at 93-4; *R v Bush* (1974-5) 5 ALR 387; *He Kaw Teh v R* (1985) 60 ALR 449 at 455.)

[171] The courts of the United States have also shown resistance to accepting that a statutory offence dispenses with the requirement of culpability. The approach of the court was summarised by Burger CJ in *United States v US Gypsum Co* 438 US 422 (1977) where he remarked:

[187] In a series of cases, this court has held that where a legislative provision imposes an obligation upon an accused to establish certain facts to avoid criminal liability it constitutes a breach of the presumption of innocence as enshrined in section 25(3)(c). (See *S v Zuma and Others* 1995 (2) SA 642 (CC); 1995 (4) BCLR 401 (CC) at para 33; *S v Bhulwana*; *S v Gwadiso* 1996 (1) SA 388 (CC) 1995 (12) BCLR 1579 (CC) at para 15; *S v Mbatha*; *S v Prinsloo* 1996 (2) SA 464 (CC); 1996 (3) BCLR 293 (CC) at para 12;

The marked-up version becomes as follows, with underlining indicating a hypertext link that has been inserted:

renewed vigour has been afforded to this approach by its recent restatement in a series of decisions by the House of Lords and the Privy Council. (See *Lim Chin Aik v R* [\[1963\] AC 160 \(PC\)](#) at 172; *R v Warner* [\[1969\] 2 AC 256 \(HL \(E\)\)](#) at 271-2; *Sweet v Parsley* [\[1969\] UKHL 1](#); [\[1970\] AC 132 \(HL \(E\)\)](#) at 163;

... See *Maher v Musson* [\[1934\] HCA 64](#); [\[1934\] 52 CLR 100](#) at 104-5; *Thomas v R* [\[1937\] HCA 83](#); [\[1937\] 59 CLR 279](#) at 287-8; *Proudman v Dayman* [\[1941\] HCA 28](#); [\[1941\] 67 CLR 536](#) at 540; *Iannella v French* [\[1968\] HCA 14](#); [\[1967-1968\] 119 CLR 84](#) at 93-4; *R v Bush* [\[1974-5\] 5 ALR 387](#); *He Kaw Teh v R* [\[1985\] HCA 43](#); [\[1985\] 60 ALR 449](#) at 455.)

[171] The courts of the United States have also shown resistance to accepting that a statutory offence dispenses with the requirement of culpability. The approach of the court was summarised by Burger CJ in *United States v US Gypsum Co* [\[1978\] USSC 150](#); [438 US 422](#) (1977) where he remarked:

[187] In a series of cases, this court has held that where a legislative provision imposes an obligation upon an accused to establish certain facts to avoid criminal liability it constitutes a breach of the presumption of innocence as enshrined in section 25(3)(c). (See *S v Zuma and Others* [\[1995\] ZACC 1](#); [1995 \(2\) SA 642 \(CC\)](#); 1995 (4) BCLR 401 (CC) at para 33; *S v Bhulwana*; *S v Gwadiso* 1996 (1) SA 388 (CC) 1995 (12) BCLR 1579 (CC) at para 15; *S v Mbatha*; *S v Prinsloo* [\[1996\] ZACC 1](#); [1996 \(2\) SA 464 \(CC\)](#); 1996 (3) BCLR 293 (CC) at para 12;

Note that (i) the marked-up text includes many more parallel citations than the source text; (ii) new links have been created to many cases, either to the source texts of the cases, or to their citation entries in the LawCite citator (where the source text is not held by a participating free access LII); and (iii) they include hypertext links to cases from South Africa, the USA, New Zealand, Australia and

the UK, demonstrating the international nature of the databases underlying LawCite.

### 7.3.2 Lawcite as 'glue'

We refer to one aspect of LawCite as 'glue': it is a means of 'glueing' documents together that are found in different databases or on different LIIs, by virtue of the citations that they share to other documents. For example, document A on BAILII and document B on PacLII may each refer to document C. If document C is located on HKLII, both A and B will link to HKLII. If document C is not found on any participating LII, both A and B will link to document C's record in LawCite.

### 7.4 The LawCite Markup Tool – a public access markup interface

The LawCite Automated Markup Tool<sup>18</sup> enables any user to mark up a whole document by insertion of a URL, or a block of pasted text. This allows anyone who wishes to publish text in electronic form to automate links to LIIs from that text. Inclusion of any known citation to a case, article, law reform report or treaty will result in a link to the text if it is available on a collaborating LII, and otherwise to the citation entry in LawCite. If the neutral citation has not been included, it will be added to the marked-up text, with a link to the free-access version of the text. Only Australian and New Zealand legislation references will be marked up, because we have not yet developed automated citation recognition for other legislation formats.

**LawCite Automated Markup Tool**

Where is the text to be processed?

Enter a URL: eg. <http://www.uni.edu.au/~user/law.html>

**OR**

Paste the text to be processed here:

Figure 6 LawCite markup tool

The effectiveness of the Markup Tool can be tested by taking the input text provided in the previous section and pasting it into the above window. The resulting output can be checked against the output example given above. It will include all the additional parallel citations, and the hypertext links. In fact, it will include more hypertext links (eg a link to [1995 \(4\) BCLR 401 \(CC\)](#)), because in the interim LawCite has acquired more citation data in the Citations database.

<sup>18</sup> LawCite Automated Markup Tool < <http://www.austlii.edu.au/LawCite/markup.html> > .

### 7.5 Alternative ranking method for search results By Citation Frequency

Most of the participating LIIs<sup>19</sup> have added 'By Citation Frequency' as an extra method of ranking search results on their systems. This ranking method depends upon the integration of the search script with the citation statistics constructed as part of the LawCite Citations database creation process.

After the default display of results 'By Relevance', selection of the 'By Citation Frequency' tab shown below will cause a display of results ranked by the number of citations of each item known to LawCite, such as:



Figure 7 Alternative search results view option: 'By Citation Frequency'

Search results on these LIIs now display (as shown above), for any case, article, law reform report or treaty in the search results list, the 'LawCite' logo which is a direct link to the citation record for that case etc in LawCite. It also displays the number of stars (1-5) indicating how frequently that item has been cited by other items. Hovering over the stars results in a display of the exact number of citations for the item.

### 7.6 Augmenting Court databases with citation data

Citation data can be further integrated into databases, for example, by the inclusion of the most cited documents for a specific database.

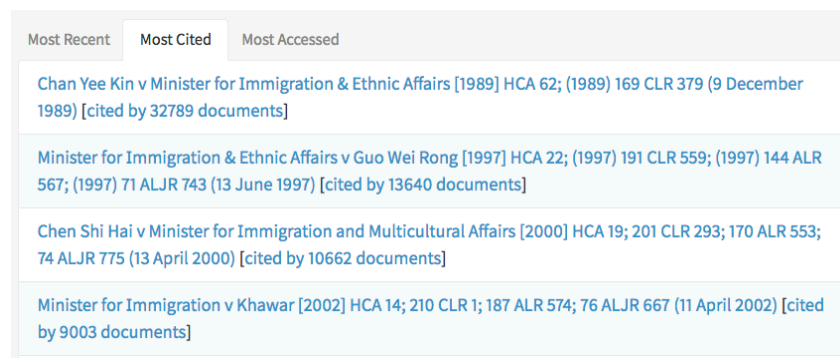


Figure 8 Most cited documents for a particular database: High Court of Australia example

### 7.7 Future applications – Research and products

The Document and Citations databases include very large data sets. These are increasingly comprehensive for some common law jurisdictions (at least for some periods). Other types of applications, both research uses and development

<sup>19</sup> This is done on AsianLII, AustLII, CommonLII, HKLII, LiberLII, LII of India, NZLII, PaCLII and WorldLII.

of products, are now possible. A few of these possible future applications are sketched in this section.

AustLII holds the full texts of almost all reported Australian case law from 1788-1960, and almost all decisions of Australian courts and tribunals since 1995. Other LIIs, including NZLII, PacLII, HKLII and BAILII, hold near-comprehensive collections of case law for differing periods. The comprehensiveness of these collections make possible certain research applications, examples of which are given below.

### 7.7.1 Informing policy and academic debates by citation data

There is often a question of what value ‘authorised’ series of reports add to a jurisdiction’s legal system. Part of that value is said to arise from the selectivity involved in authorised reports: they select for reporting only those decisions which expert editors judge to be of future precedential value. If such judgments/predictions are of value, then arguably the selected cases from a particular year should be cited more often in future cases than other cases decided by the same court in the same year. Frequency of future citation is not the only measure of precedential value (for example, where a case determines a legal point conclusively), but it is certainly a very relevant one (see further van Opijnen 2013).

The value of such selectivity can be tested by examining the history of how often the selected cases in the authorised series are cited by subsequent cases (or other documents), compared with citations of other cases by the same court which have been ‘left out’ of the authorised series. Because AustLII holds almost all Australian decisions since 1995, LawCite holds comprehensive data on Australian citations of Australian cases since then. Tables generated from the LawCite citations database show, for example, that, from all known citations of New South Wales Supreme Court decisions, the percentage of cited decisions being to those included in the Court’s authorised reports has declined from 100% in 1995 to 36% in 2005 and 18% in 2013. Such initial statistics, while suggestive, require further research before conclusions can be drawn.

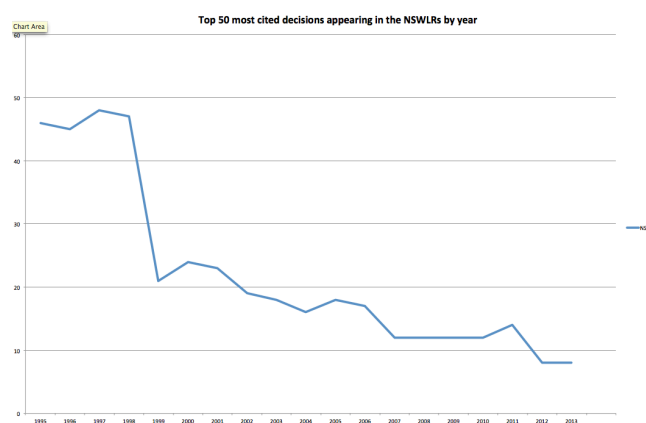


Figure 9 Top 50 most cited decisions in the NSWLRs by year

Some important questions of post-colonial jurisprudence are now capable of being answered. For example, PacLII now holds most case-law from Pacific Island countries since at least 2000. For example, by using LawCite’s Citations

database, it may be possible (assuming case citations are used) to calculate what percentage of decisions cited by any courts in Pacific Island jurisdictions are to decisions made in other Pacific Island jurisdictions,<sup>20</sup> compared with decisions of courts outside the Pacific (eg those in the UK, Australia and New Zealand). It would be valuable to know whether these percentages are changing over time, and for scholars of Pacific Islands law to consider the reasons.

On the other side of the coin, it would be interesting to know whether, since their growing international availability (thanks largely to PacLII) since 2000, decisions by courts in Pacific Island countries have been cited increasingly by courts in countries outside the Pacific, and from where have such citations occurred? Such answers would not be comprehensive, because LawCite does not hold comprehensive citation data for many jurisdictions outside the Pacific. However, for example, a substantially comprehensive study could include courts in the UK, Australia, New Zealand, Hong Kong and South Africa since 2000.

Similar research on citation of articles from law journals by Australian courts and tribunals can also be given comprehensive answers (at least since 1995) to questions of importance to some audiences. Answerable questions include 'which law journals are most frequently cited by Australian courts or tribunals (both overseas journals and Australian ones), and is this changing over time?'

### **7.7.2 Data visualisation of citation flows**

The results of the hypothetical Pacific Islands citation research mentioned above could be presented by data visualisation tools. Visualisation facilities are now available to LawCite researchers and tools are being developed to utilise these facilities to represent visually the citation trends and patterns.<sup>21</sup> The flows of citation to and from Pacific Island jurisdictions may make more sense as visualisations.

Looking ahead, it is possible to imagine LawCite data providing near-comprehensive coverage of citations in all common law countries over time. If so, the notion of the common law as an international, interconnected, legal system where the connections are provided primarily by citations, could be interrogated. In all probability, the best way to understand such a vast body of interconnected data would be through data visualisation tools.

### **7.7.3 Contextual ranking**

These hypertext links created through citation tables will also be used to build a new form of relevance ranking of search results suitable for legal materials. Traditional methods of relevance ranking (as AustLII currently uses) based on word-occurrence density and position are effective but require improvement to take account of legally specific features of text. Approaches utilising the number of links to a document such as implemented by Google (Brin and Page, 1998) may be an indicator of significance, but taken alone are too coarse a measure, because 'popular' cases which are often cited for general principles will be

---

<sup>20</sup> Distinctions could be made as to whether their own jurisdiction should or should not be included.

<sup>21</sup> For example, the UTS Data Arena is a 360-degree interactive data visualisation facility which utilises six 3D-stereo video projectors to create a seamless three-dimensional panorama.

unduly highly ranked whenever found. A more precise approach requires consideration of the relative frequency of links from documents in the retrieved set: we call this 'contextual ranking'. The aim of this approach will be to rank within a set of results that have known relevance. Preliminary work shows that, whilst computationally expensive, this might yield far better ranking results than any existing legal search retrieval system. This would also give much better ranking of the 'noteup' results. This difficult research will draw on work done in comparing the effectiveness of using 'local' versus 'global' link information in improving document retrieval (Calado et al, 2003; Kleinberg, 1998; van Opijnen 2013).

## 8. Further development of LawCite tools

Separate from the possible applications of the LawCite tools, further development of the tools themselves are needed. For example, there is considerable potential for LawCite to better handle international citations, in particular, for conflicting series names.

Also, the production of updates to Citations and Documents databases, and thus to the Citator itself can be improved. It is currently too slow and resource-intensive, because of the vast numbers of documents which must be data mined, and the complexity of the unmining tasks. Ideally, updates should occur daily, but storage and processing resources available to AustLII until now have only enabled updates to be generated monthly (or thereabouts). This will be resolved in 2016 with the acquisition of new hardware, after which it is expected updates will occur at least weekly (and ideally, daily).

## References

- Borkowski, C (1969) 'Structure, Effectiveness and Uses of the Citation Identifier', Stockholm: *International Conference on Computational Linguistics*. Coling. Association for Computing Machinery.
- Borodin A, Roberts G O, Rosenthal J S, and Tsaparas P (2005), 'Link analysis ranking: algorithms, theory, and experiments', *ACM Transactions on Internet Technology (TOIT)*, Vol 5 Issue 1, pp 231-297.
- Brin S and Page L (1998) 'The anatomy of a large-scale hypertextual Web search engine'. In *Proceedings of the 7<sup>th</sup> International World Wide Web Conference*, pp107-117, Brisbane, Australia
- Calado P, Ribeiro-Neto B, Ziviani N, Moura E, and Silva I (2003), 'Local versus global link information in the Web', *ACM Transactions on Internet Technology (TOIT)*, Vol 21 Issue 1, pp 42-63.
- Declaration (2002) Declaration on Free Access to Law, Montreal 2002 (as amended) <<http://falm.info/declaration/>>.
- Giles C L, Bollacker K, and Lawrence S (1998) 'CiteSeer: An Automatic Citation Indexing System,' *Digital Libraries 98: Third ACM Conf. Digital Libraries*, ACM Press, New York, pp 89-98.
- Greenleaf G, Chung P, and Mowbray A (2007) 'Emerging global networks for free access to law: WorldLII's strategies' (2007) 4:4 SCRIPT-ed 319 at <<http://www.law.ed.ac.uk/ahrc/script-ed/vol4-4/greenleaf.asp>>.



- Kleinberg J M (1998), 'Authoritative sources in a hyperlinked environment'. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pp 668-677, San Francisco, California, United States.
- Lawrence S, Giles C L, and Bollacker K (1999) 'Digital Libraries and Autonomous Citation Indexing' *IEEE Computer*, 32, 6, 67-71, 1999  
<<http://clgiles.ist.psu.edu/papers/IEEE.Computer.DL-ACI.pdf>>.
- Poulin D, Pare E, and Mokanov I (2005) 'RefLex – Bridging Open Access with a Legacy Legal Information System'. In *Law via the Internet 2005*, Port Vila, November, 2005
- Nigam K, McCallum A, and Ungar L H (2000) 'Efficient clustering of high-dimensional data sets with application to reference matching'. In *Conference on Knowledge Discovery in Data - Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178. ACM, 2000.
- Mowbray A, Austin D, and Chung P (2000) 'Scalability of Web Resources for Law: AustLII's Technical Roadmap: Past, Present and Future', 2000 (1) *The Journal of Information, Law and Technology* (JILT) <<http://www.law.warwick.ac.uk/jilt/00-1/austin.html>>.
- Mowbray A, Chung P and Greenleaf G (2009) 'Free-access case law enhancements for Australian law' in *Free Access Quality of Information Effectiveness of Rights*, Peruginelli and Ragona (Eds), European Press Academic Publishing, Florence Italy.
- Mowbray A, Greenleaf G and Chung P (2000) 'A Uniform Approach for Vendor and Media Neutral Citation – the Australian Experience' Citations Workshop, University of Edinburgh, Scotland, 2000.
- Olsson, The Honourable Justice L T (2nd Ed., 1999) *Guide to Uniform Production of Judgments*, AIJA.
- Rogers I 'The Google Pagerank Algorithm and How It Works' IPR Computing Ltd,  
<<http://www.iprcom.com/papers/pagerank/>>.
- Rugh J and Lennen J (2003) 'Using Fuzzy Logic to Create Links: Resolving References to Court Decisions' XML Conference and Exposition, 2003  
<[http://www.idealliance.org/papers/dx\\_xml03/papers/05-05-03/05-05-03.html](http://www.idealliance.org/papers/dx_xml03/papers/05-05-03/05-05-03.html)>.
- Salton G (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison-Wesley Longman Publishing).
- van Opijnen M (2013) 'A Model for Automated Rating of Case Law'. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (ICAIL)*, 10-14 June, 2013, Rome, Italy <<http://dl.acm.org/citation.cfm?id=2514617>>.
- van Opijnen M, Verwer N, and Meijer J (2015) 'Beyond the Experiment: The Extendable Legal Link Extractor'. In *Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, held in conjunction with the *2015 International Conference on Artificial Intelligence and Law (ICAIL)*, 8-12 June, 2015, San Diego, CA, USA <[http://ssrn.com/abstract\\_id=2626521](http://ssrn.com/abstract_id=2626521)>.
- Wittfoth A, Chung P, Mowbray A, and Greenleaf G (2003) 'Can One Size Fit All?: - AustLII's Point-in-Time Legislation Project', *5th Law via Internet Conference*, Sydney, 2003.
- Wittfoth A, Chung P, Greenleaf G, and Mowbray A (2005), 'AustLII's Point-in-Time legislation system: A generic PiT system for presenting legislation', Launch of the Point-in-Time legislation system, 7 April 2005.
- Wong S K M, Ziarko W, Raghavan V V, and Wong P C N (1987) 'On Modeling of Information Retrieval Concepts in Vector Spaces', *ACM Transactions on Database Systems*, Vol 12, No 2, pp 299-321.

## Appendix 1 – Series database summary

(Generated October 2011)

| Country | Cases | Journals | Total |
|---------|-------|----------|-------|
| AD      | 4     |          | 4     |
| AL      | 5     |          | 5     |
| AR      | 10    | 1        | 11    |
| AS      | 3     |          | 3     |
| AT      | 45    | 3        | 48    |
| AU      | 665   | 458      | 1123  |
| BA      | 4     |          | 4     |
| BB      | 3     | 2        | 5     |
| BD      | 5     | 1        | 6     |
| BE      | 47    | 20       | 67    |
| BF      | 1     | 1        | 2     |
| BG      | 11    |          | 11    |
| BI      |       | 1        | 1     |
| BM      | 1     |          | 1     |
| BN      | 1     |          | 1     |
| BR      | 6     | 7        | 13    |
| BS      | 2     |          | 2     |
| BW      | 5     | 1        | 6     |
| CA      | 786   | 186      | 972   |
| CD      |       | 1        | 1     |
| CG      | 3     | 1        | 4     |
| CH      | 2     | 6        | 8     |
| CI      |       | 2        | 2     |
| CK      | 5     |          | 5     |
| CL      | 2     | 4        | 6     |
| CM      | 4     |          | 4     |
| CN      | 25    | 16       | 41    |
| CO      | 1     | 2        | 3     |
| CR      | 1     |          | 1     |
| CU      | 3     | 2        | 5     |
| CY      | 3     | 1        | 4     |
| CZ      | 8     | 2        | 10    |
| DE      | 12    | 30       | 42    |
| DK      | 11    | 2        | 13    |
| DZ      | 1     | 3        | 4     |
| EC      | 2     |          | 2     |
| EE      | 2     | 1        | 3     |
| EG      | 4     |          | 4     |

| Country | Cases | Journals | Total |
|---------|-------|----------|-------|
| ES      | 21    | 19       | 40    |
| ET      | 1     |          | 1     |
| EU      | 48    | 35       | 83    |
| EW      | 1     |          | 1     |
| FI      | 11    | 1        | 12    |
| FJ      | 9     |          | 9     |
| FM      | 7     |          | 7     |
| FR      | 1     | 6        | 7     |
| GE      |       | 1        | 1     |
| GG      |       | 1        | 1     |
| GH      | 8     | 2        | 10    |
| GI      | 1     |          | 1     |
| GL      | 1     |          | 1     |
| GR      | 13    |          | 13    |
| GU      | 5     |          | 5     |
| GY      | 13    | 1        | 14    |
| HK      | 29    | 12       | 41    |
| HR      | 8     | 2        | 10    |
| HT      | 1     |          | 1     |
| HU      | 8     | 3        | 11    |
| ID      | 7     | 3        | 10    |
| IE      | 107   | 49       | 156   |
| IL      | 4     | 10       | 14    |
| IM      | 3     |          | 3     |
| IN      | 417   | 98       | 515   |
| INT     | 31    | 1        | 32    |
| IS      | 3     |          | 3     |
| IT      | 266   | 10       | 276   |
| JE      | 1     |          | 1     |
| JM      | 13    | 2        | 15    |
| JO      | 1     |          | 1     |
| JP      | 75    | 20       | 95    |
| KE      | 14    |          | 14    |
| KH      | 3     | 1        | 4     |
| KI      | 5     |          | 5     |
| KR      | 10    | 10       | 20    |
| LA      | 1     | 1        | 2     |
| LB      | 2     |          | 2     |

| Country | Cases | Journals | Total |
|---------|-------|----------|-------|
| LI      | 3     |          | 3     |
| LK      | 65    | 7        | 72    |
| LR      | 2     | 1        | 3     |
| LS      | 4     |          | 4     |
| LT      | 2     | 1        | 3     |
| LU      | 6     | 1        | 7     |
| LV      | 1     |          | 1     |
| LY      | 1     |          | 1     |
| MA      | 2     | 1        | 3     |
| MC      | 1     |          | 1     |
| MG      | 3     |          | 3     |
| MH      | 3     |          | 3     |
| MK      | 1     |          | 1     |
| MM      | 23    | 2        | 25    |
| MO      | 6     |          | 6     |
| MP      | 3     |          | 3     |
| MT      | 1     | 1        | 2     |
| MU      | 7     |          | 7     |
| MW      | 8     | 1        | 9     |
| MX      | 8     | 3        | 11    |
| MY      | 36    | 10       | 46    |
| MZ      | 1     |          | 1     |
| NA      | 3     |          | 3     |
| NG      | 36    | 4        | 40    |
| NIE     | 1     |          | 1     |
| NL      |       | 64       | 64    |
| NO      | 6     |          | 6     |
| NP      |       | 1        | 1     |
| NR      | 3     |          | 3     |
| NU      | 2     |          | 2     |
| NZ      | 151   | 65       | 216   |
| PA      | 1     |          | 1     |
| PE      | 2     | 4        | 6     |
| PG      | 13    |          | 13    |
| PH      | 10    | 7        | 17    |
| PK      | 38    | 8        | 46    |
| PL      |       | 1        | 1     |
| PN      | 3     |          | 3     |
| PO      |       | 1        | 1     |
| PR      | 10    | 4        | 14    |
| PS      | 1     |          | 1     |
| PT      | 13    | 3        | 16    |

| Country      | Cases       | Journals    | Total        |
|--------------|-------------|-------------|--------------|
| PW           | 1           |             | 1            |
| PY           | 1           |             | 1            |
| RO           | 9           | 7           | 16           |
| RS           | 2           | 3           | 5            |
| RW           |             | 2           | 2            |
| SB           | 5           |             | 5            |
| SC           | 2           |             | 2            |
| SD           | 3           |             | 3            |
| SE           |             | 1           | 1            |
| SG           | 11          | 10          | 21           |
| SI           | 1           | 2           | 3            |
| SL           | 5           | 1           | 6            |
| SN           | 1           | 2           | 3            |
| SV           | 1           |             | 1            |
| SZ           | 5           |             | 5            |
| TH           | 5           | 2           | 7            |
| TL           | 1           |             | 1            |
| TN           | 1           | 1           | 2            |
| TO           | 5           |             | 5            |
| TR           | 7           |             | 7            |
| TT           | 4           |             | 4            |
| TV           | 2           |             | 2            |
| TW           | 2           | 3           | 5            |
| TZ           | 5           | 2           | 7            |
| UA           | 2           |             | 2            |
| UG           | 10          | 1           | 11           |
| UK           | 1330        | 425         | 1755         |
| US           | 1967        | 1888        | 3855         |
| UY           | 4           | 1           | 5            |
| UZ           | 1           |             | 1            |
| VA           | 1           |             | 1            |
| VE           | 2           | 2           | 4            |
| VI           | 2           |             | 2            |
| VN           | 2           | 1           | 3            |
| VU           | 8           | 1           | 9            |
| WS           | 8           |             | 8            |
| YE           | 1           |             | 1            |
| ZA           | 117         | 31          | 148          |
| ZM           | 6           | 1           | 7            |
| ZW           | 10          | 3           | 13           |
| <b>Total</b> | <b>6872</b> | <b>3622</b> | <b>10494</b> |