

Important Note: Please read the methodology notes for Legislative Data in Interim Report of the Financial Services Inquiry if you wish to understand the full limitations of the data. The key limitation is the inconsistent quality of the HTML for each piece of legislation that is obtained from Federal Register of Legislation, particularly in older legislation, treaties, or intergovernmental agreements. Variables that rely on analysing HTML may be an undercount. Some older pieces of legislation or treaties and intergovernmental agreements do not use markup at all in the HTML for certain textual data, such as chapter, section, and subsection numbers.

Important concepts – Legislation

The following two definitions are based on the Federal Register of Legislation Glossary. For more information, see <https://www.legislation.gov.au/content/whatisit>. The Federal Register of Legislation also publishes a guide to the typical structure of a Commonwealth Act: <https://www.legislation.gov.au/file/StructureOfAnAct>.

As made legislation is legislation in the form in which it was originally made. The text may not reflect in force as it will not include any later amendments that may have been made.

A *compilation* is a version of a law that shows the text of the law as amended at a particular point in time.

Important concepts – Process of analysis

Stopwords are words that are removed for certain types of text analysis, such as determining the number of unique words in a text (ie its vocabulary). The ALRC uses R's Quanteda package, which includes 172 stopwords.

Excluded structural elements refers to subsection, paragraph, subparagraph, and sub-subparagraph numbers and lettering that appear at the beginning of lines in legislation (eg (1), (d), (iv), (A)). The ARLC removes these for certain types of text analysis.

Legislation Name	Name of the legislation as scraped. Based on name of the legislation to which the text relates. For example, the <i>Competition and Consumer Act 2010</i> was previously called the <i>Trade Practices Act 1974</i> . The name will depend on the specific text (as at a particular date) being scraped.
FRL ID	The unique identification number given to the piece of legislation by the Federal Register of Legislation.
Administrator	The administrator of the legislation as listed under the 'Details' tab of the legislation page – see 'Administered by:' under this tab. Legislation can have more than one administrator.
Date of Registration	Date of registration for the legislation as listed under the 'Details' tab of the legislation page – see 'Registered' under this tab.
Date of Assent	Date of Royal Assent for the legislation as listed under the 'Details' tab of the legislation page – see 'Date of Assent' under this tab. Only applicable to Acts of Parliament as made. Compilations have a 'Start Date'.
Start Date	Date from which a legislation compilation takes effect as listed under the 'Details' tab of the legislation page – see 'Start Date' under this tab. Only applicable to compilations.

End Date	Date on which a legislation compilation ceases to have effect as listed under the ‘Details’ tab of the legislation page – see ‘End Date’ under this tab. Only applicable to compilations.
Date of Repeal	Date on which an as made piece of legislation was repealed. Only applicable to legislation as made. Compilations have an ‘End Date’.
Legislation Number/Year	The first non-italicised paragraph description that appears under the ‘Details’ tab of the legislation page.
Year	The year extracted from the ‘Legislation Number/Year’ variable
Years in operation	The product of the ‘Year’ variable subtracted from 2021.
Amendments per year	‘Amendments to Act’ variable divided by ‘Years in operation’ variable.
Summary	The second italicised paragraph description that appears under the ‘Details’ tab of the legislation page.
Legislation Pages	The number of pages of the legislation as appears on the ‘Download’ page for the legislation.
Number of Volumes	The number of volumes of the legislation as appears on the ‘Download’ page for the legislation.
Amendments to Act	The number of amendments to the legislation based on the ‘Act Amendments’ tab that appears on the ‘Principal + Amendments’ page. This page is accessed through the ‘Series’ page of the legislation.
Modifications to Act	The number of amendments to the legislation based on the ‘Modified By / Affected By’ page. This page is accessed through the ‘Series’ page of the legislation.
Instruments Enabled	The number of amendments to the legislation based on the ‘Enables’ page. This page is accessed through the ‘Series’ page of the legislation.
Related Bills	The number of amendments to the legislation based on the ‘Related Bills’ page. This page is accessed through the ‘Series’ page of the legislation.
Word count	Counts the number of ‘tokens’ using R’s Quanteda package, splitting the document using ‘fasterword’. Does not include endnotes, table of contents, and excluded structural elements. Does not split hyphens.
Entropy	<p>Calculated using the below equation from Patrick McLaughlin et al, ‘Is Dodd-Frank the Biggest Law Ever?’ (2021) 7(1) Journal of Financial Regulation 149, 170. Does not include endnotes, table of contents, and excluded structural elements. Does not split hyphens. Also excludes stopwords, numbers, and alphanumeric words (eg 601AKC).</p> $H(D) = -\sum_{w \in W_D} p_w \log_2(p w),$ <p>‘where D is a document, H(D) is the Shannon entropy of document D, WD is the set of unique words occurring in document D, and pw is the probability of encountering one of these words at a random point in the text—that is, the frequency of that word as a percentage of the total word count.’</p>
Number of unique substantive words	Counts the number of unique ‘tokens’ using R’s Quanteda package. Does not include endnotes, table of contents, and excluded structural elements. Does split hyphens. Also excludes stopwords, numbers, and alphanumeric words (eg 601AKC).

Average substantive word length	Average word length in a text. Does not include endnotes, table of contents, and excluded structural elements. Does split hyphens. Also excludes stopwords, numbers, and alphanumeric words (eg 601AKC).
Unique word stems for substantive words	The number of unique stem words that appear in a text. For example, the stemmed word of ‘provider’ and ‘provided’ is ‘provid’. Does not include endnotes, table of contents, and excluded structural elements. Does split hyphens. Also excludes stopwords, numbers, and alphanumeric words (eg 601AKC).
Readability	Flesch-Kinkaid score - the lower the score the less readable the text. Has limitations in legislation because headings do not use end of sentence punctuation. This means the sentence length can be overestimated, which makes the text appear less readable.
Average section length	The word count divided by number of sections. See description of ‘Word count’ for exclusions from word count.
Chapters	Number of Chapters marked-up in the HTML. Some Chapters are marked-up without content, and these are removed, as are ‘placeholder’ Chapters that appear in the HTML. Duplicate Chapters are removed based on their full name. For example, ‘Chapter 7—Financial services and markets’ appears in both Volumes 4 and 5 of the <i>Corporations Act 2001</i> . This is only counted once. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
Schedules	Number of Schedules marked-up in the HTML. Some Schedules are marked-up without content, and these are removed, as are ‘placeholder’ Schedules that appear in the HTML. Duplicate Schedules are removed based on their full name. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
Parts	Number of Parts marked-up in the HTML. Some Parts are marked-up without content, and these are removed, as are ‘placeholder’ Parts that appear in the HTML. Duplicate Parts are removed based on their full name. Duplicates are therefore only counted once. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
Divisions	Number of Divisions marked-up in the HTML. Some Divisions are marked-up without content, and these are removed, as are ‘placeholder’ Divisions that appear in the HTML. Duplicate Divisions are removed based on their full name. Duplicates are therefore only counted once. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
Subdivisions	Number of Subdivisions marked-up in the HTML. Some Subdivisions are marked-up without content, and these are removed, as are ‘placeholder’ Subdivisions that appear in the HTML. Duplicate Subdivisions are removed based on their full name. Duplicates are therefore only counted once. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
Sections	Number of Sections marked-up in the HTML (eg 423A, 732). Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
Subsections	Number of Subsections marked-up in the HTML (eg (1), (12)).

Paragraphs	Number of Paragraphs marked-up in the HTML (eg (a), (aa)).
Subparagraphs	Number of Subparagraphs marked-up in the HTML (eg (i), (iv)).
Sub-subparagraphs	Number of Sub-subparagraphs marked-up in the HTML (eg (A), (B)).
Notes	Number of Notes marked-up in the HTML. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
Elements above section	Sum of Chapters + Schedules + Parts + Divisions + Subdivisions
Elements section and below	Sum of Sections + Subsections + Paragraphs + Subparagraphs + Sub-subparagraphs
Amending schedules	Number of Amending Schedules marked-up in the HTML (ie Schedules of a piece of legislation that contain amendments to another piece of legislation). This variable can help identify legislation that contains amendments to other pieces of legislation.
Amending parts	Number of Amending Parts marked-up in the HTML (ie Parts of a piece of legislation that contain amendments to another piece of legislation). This variable can help identify legislation that contains amendments to other pieces of legislation.
Items	Number of Items marked-up in the HTML (ie 'ItemHead'). Items generally contain amendments to another piece of legislation. This variable can help identify legislation that contains amendments to other pieces of legislation.
Defined terms	Number of defined terms marked-up in the HTML. Because defined terms are accompanied by a definition, this is also a count of definitions.
Unique defined terms	Number of defined terms with duplicated defined terms removed (eg if 'property' is defined three times, it is only counted once here).
Number of potential uses of defined terms	The number of times a potentially defined term is used in the legislation. Determined using a list of all terms marked up in the HTML as defined terms in the piece of legislation. The use of a term is not counted where it appears in the use of another defined term (to avoid duplication). For example, 'financial product advice', a defined term, is counted and the use of 'financial product', another defined term, is not counted when it appears in that defined term. Terms are 'potentially' defined because not all definitions apply for all provisions in a piece of legislation, so a term may be used in an undefined sense even if defined for other provisions.
Number of words potentially defined	The number of words that are potentially defined in an Act, determined using the same approach for the 'Number of potential uses of defined terms' variable but counting words comprising the terms rather than uses of the terms. For example, while 'financial product advice' will only count as one use of a defined term it will count for three words that are potentially defined. Words are 'potentially' defined because not all definitions apply for all provisions in a piece of legislation, so a word may be used in an undefined sense even if defined for other provisions.
Potentially defined words used per 100 words	The 'Number of words potentially defined' divided by the 'Word count'.
Bold and italicised terms	Number of bold and italicised terms marked-up in the HTML.
Unique bold and italicised terms	Number of bold and italicised terms with duplicated defined terms removed (eg if 'property' is defined three times, it is only counted once here).

Number of potential uses of bold and italicised terms	The number of times a term is used in the legislation that is potentially affected by a bold and italicised term. Determined using a list of all terms marked up in the HTML as bold and italicised terms in the piece of legislation. The use of a term is not counted where it appears in the use of another bold and italicised term (to avoid duplication). For example, ‘financial product advice’, a bold and italicised term, is counted and the use of ‘financial product’, another bold and italicised term, is not counted when it appears in that bold and italicised term. Terms are ‘potentially’ used because not all bold and italicised terms apply for all provisions in a piece of legislation, so a term may be used in an undefined or untagged sense even if defined or tagged for other provisions.
Number of words potentially bold and italicised	The number of words that are potentially affected by a bold and italicised term in the legislation, determined using the same approach for the ‘Number of potential uses of bold and italicised terms’ variable but counting words comprising the terms rather than uses of the terms. For example, while ‘financial product advice’ will only count as one use of a bold and italicised term it will count for three words that are potentially affected by a bold and italicised term. Words are ‘potentially’ bold and italicised because not all bold and italicised terms apply for all provisions in a piece of legislation, so a word may be used in an undefined or untagged sense even if defined or tagged for other provisions.
Potentially bold and italicised words used per 100 words	The ‘Number of words potentially bold and italicised’ divided by the ‘Word count’.
Number of potential uses of defined terms (both regulations and Act terms)	Only applicable to regulations. The same methodology as ‘Number of potential uses of defined terms’ variable but adding all terms defined in the Act that authorises the regulations to the terms that are defined in the regulations.
Number of words potentially defined (both regulations and Act terms)	Only applicable to regulations. The same methodology as ‘Number of words potentially defined’ variable but adding all terms defined in the Act that authorises the regulations to the terms that are defined in the regulations.
Potentially defined words used per 100 words (both regulations and Act terms)	The ‘Number of words potentially defined (both regulations and Act terms)’ divided by the ‘Word count’.
Act cross-references	Counts the number of references to ‘^Act\$’ that appear in the text of the legislation. The code then deletes results that are immediately preceded by any of the following terms: ‘[Tt]his’, ‘An’, ‘)’, ‘[Tt]hat’, ‘[Tt]he’, or ‘[Aa]pplication’. The ALRC identified these terms did not indicate an external cross-reference.
Cross references to other Acts per 100 words	‘Act cross-references’ divided by the ‘Word count’.
Internal cross-references – Sections	Counts the number of references to ‘^section.*’, ‘^subsection.*’, ‘^subparagraph.*’, ‘^paragraph.*’. The code then deletes results that are immediately preceded by ‘this’ or followed by ‘of’. The ALRC identified these terms did not indicate an internal cross-reference.
Cross references to other sections per 100 words	‘Internal cross-references – Sections’ divided by the ‘Word count’.
Cross-references – Regulations	Only applicable to Regulations. Counts the number of references to ‘^regulation\$’, ‘^subregulation.*’. The code then deletes results that are immediately preceded by ‘this’ or followed by ‘of’. ALRC identified these terms did not indicate an internal cross-reference.

Cross references to Regulations per 100 words	‘Cross-references – Regulations’ divided by the ‘Word count’.
Conditional Statements	Counts the number of the following that appear in the text of the legislation: ‘^if\$’, ‘^except\$’, ‘^but\$’, ‘^provided\$’, ‘^when\$’, ‘^where\$’, ‘^whenever\$’, ‘^unless\$’, ‘^notwithstanding\$’
Conditional statements per 100 words	‘Conditional Statements’ divided by the ‘Word count’.
Obligations	Counts the number of the following that appear in the text of the legislation: ‘^must\$’, ‘^shall\$’, ‘^may not’, ‘^prohibited’, ‘^required’, ‘^may only’, ‘^cannot be’
Obligations per 100 words	‘Obligations’ divided by the ‘Word count’.
Offences	Counts the number of the following that appear in the text of the legislation: ^offence\$
Offences per 100 words	‘Offences’ divided by the ‘Word count’.
Reasonableness	Counts the number of the following that appear in the text of the legislation: ^reasonabl.*
Reasonableness per 100 words	‘Reasonableness’ divided by the ‘Word count’.
Modifications	Counts the number of the following that appear in the text of the legislation: ^omit.*, ‘^insert\$’, ‘^substitute\$’
Modifications per 100 words	‘Modifications’ divided by the ‘Word count’.
Contravene	Counts the number of the following that appear in the text of the legislation: ^contravene.*
Contravene per 100 words	‘Contravene’ divided by the ‘Word count’.
Discretions	Counts the number of the following that appear in the text of the legislation: ‘Minister may’, ‘ASIC may’, ‘ACCC may’, ‘RBA may’, ‘APRA may’
Discretions per 100 words	‘Discretions’ divided by the ‘Word count’.
Legislative instruments	Counts the number of the following that appear in the text of the legislation: ‘legislative instrument’
Legislative instruments per 100 words	‘Legislative instruments’ divided by the ‘Word count’.
Regulations	Counts the number of the following that appear in the text of the legislation: ^regulations\$
Regulations per 100 words	‘Regulations’ divided by the ‘Word count’.
Strict liability	Counts the number of the following that appear in the text of the legislation: strict liability
Strict liability per 100 words	‘Strict liability’ divided by the ‘Word count’.
Civil liability	Counts the number of the following that appear in the text of the legislation: civil penalt.*
Civil liability per 100 words	‘Civil liability’ divided by the ‘Word count’.
Misleading	Counts the number of the following that appear in the text of the legislation: mislead.*
Misleading per 100 words	‘Misleading’ divided by the ‘Word count’.
Unconscionable	Counts the number of the following that appear in the text of the legislation: unconsciona.*
Unconscionable per 100 words	‘Unconscionable’ divided by the ‘Word count’.
Dishonesty	Counts the number of the following that appear in the text of the legislation: dishonest.*
Dishonesty per 100 words	‘Dishonesty’ divided by the ‘Word count’.
Honesty	Counts the number of the following that appear in the text of the legislation: ^honest.*

Honesty per 100 words	'Honesty' divided by the 'Word count'.
Good faith	Counts the number of the following that appear in the text of the legislation: good faith
Good faith per 100 words	'Good faith' divided by the 'Word count'.
Unfair	Counts the number of the following that appear in the text of the legislation: ^unfair.*
Unfair per 100 words	'Unfair' divided by the 'Word count'.
Fair	Counts the number of the following that appear in the text of the legislation: ^fair.*
Fair per 100 words	'Fair' divided by the 'Word count'.
Mistake	Counts the number of the following that appear in the text of the legislation: ^mistake.*
Mistake per 100 words	'Mistake' divided by the 'Word count'.
Unjust	Counts the number of the following that appear in the text of the legislation: ^unjust.*
Unjust per 100 words	'Unjust' divided by the 'Word count'.