Using InCites Responsibly: A Guide to Interpretation and Good Practice

Andrew Gray (UCL) and Robyn Price (Imperial College London)

https://doi.org/10.25561/75946

Introduction

This guide has been created by bibliometric practitioners to support other users of InCites and promote a community of informed and responsible use. The recommendations in this document may be more suited to other academic sector users, but the authors hope that other users may also benefit from the suggestions.

The guide aims to provide plain-English definitions, key strengths and weaknesses and some practical application tips for some of the most commonly-used indicators available in InCites. The indicator definitions are followed by explanations of the data that powers InCites, attempting to educate users on where the data comes from and how the choices made in selecting and filtering data will impact on final results. Also in this document are a comparative table to highlight differences between indicators in InCites and SciVal, another commonly used bibliometric analytic programme, and instructions on how to run group reports.

All of the advice in this document is underpinned by a belief in the need to use InCites in a way that respects the limitations of indicators as quantitative assessors of research outputs.

Both of the authors are members of signatory institutions of DORA, the San Francisco Declaration on Research Assessment.[1] A summary of advice to using indicators and bibliometric data responsibly is available on pages 4-5 and should be referred to throughout.

Readers are also recommended to refer to the official InCites Indicators Handbook produced by Clarivate Analytics.[2]

The guide was written with complete editorial independence from Clarivate Analytics, the owners of InCites. Clarivate Analytics supported the authors of this document with checking for factual accuracy only.

Acknowledgements

The structure and some indicator interpretations have been adapted from a similar work led by Dr Ian Rowlands based on SciVal.[3]

Many thanks to fellow bibliometric practitioners who provided feedback on various drafts of this document; Samuel Cole (University of Warwick), Ian Rowlands (King's College London) and Jane Falconer (London School of Hygiene and Tropical Medicine).

Thanks also to Lea Lipitakis and Rachel Mangan, both of Clarivate Analytics, who offered their expertise of the product and checked for factual accuracy.

[1] https://sfdora.org/

[2] http://help.prod-incites.com/inCites2Live/indicatorsGroup/aboutHandbook.html

[3] The SciVal guide is not available at time of writing. Email Robyn.Price@imperial.ac.uk for information or check the Bibliomagician blog or LIS Bibliometrics discussion list (see page 33 for details) for updates.

Good practice for using quantitative indicators

The below advice should be considered before using InCites or any bibliometric indicators.

The principles of the Leiden Manifesto[4] and DORA[5] should be referred to before undertaking any quantitative bibliometric analysis. Users in an institution with a responsible metrics policy or that is a signatory of DORA will be expected to uphold these standards.

We would recommend that you always try to provide multiple indicators, rather than standardising on just one. For example, a good combination might be to report CNCI or average percentile, which represent overall citation levels of all publications, alongside the share of documents in the top 1% or the share of ESI Highly Cited Papers, which concentrate on the "top papers". Giving multiple indicators helps reduce the effect of outlier papers affecting the calculations, while also allowing different aspects of research quality to be assessed.

Essential advice for selecting and interpreting indicators

- Use at least two indicators in any metric-based assessment.
- Avoid using journal-based indicators (such as the impact factor or journal rank quartiles) to assess the quality of individual papers.
- Citation-based indicators are time-dependent and thus must only be used to compare articles of similar ages. Likewise, citation-based indicators are discipline specific and should not be used to compare articles of different subject disciplines.
- Using discipline filters (Schemas or Research Areas) might not give a full picture of multidisciplinary research or work in newly-emerging disciplines.
- Any indicator calculated by a simple average will be susceptible to being skewed by outliers. Check for outliers in your dataset and consider whether they might skew the indicators you wish to use.
- Accept that the pre-set InCites indicators might not, in every case, match your analysis need. If there are not suitable indicators, users could consider exporting data for analysis in a different programme. Similarly, avoid selecting indicators on the basis that they may fit the conclusions you are expecting to draw.
- Avoid false precision – while InCites quotes many indicators to two decimal places, this can give a false sense of accuracy, and very small differences between groups or over time are unlikely to be meaningful. Consider rounding values when interpreting them.
- When presenting the data, include the date of either the most recent InCites update and/or the date of which you collected the data, alongside any filters applied to the source data. In particular, remember to note whether you included ESCI papers or not.

Sample sizes

Normalised indicators may be less reliable when they are calculated on the basis of a small group of papers, as they are more susceptible to being skewed by outliers. For most averaged indicators, such as the average CNCI or average percentile, a commonly used minimum is around 50 papers. For those that show the share of papers passing a particular threshold, such as papers in the top 1% or ESI Highly Cited papers, a higher threshold might be more appropriate; for papers in the top 1%, for

[4] Hicks, D.; Wouters, P.; Waltman, L.; de Rijcke, S.; Rafols, I. Bibliometrics: The Leiden Manifesto for research metrics. *Nature* (2015) 429-431 doi:10.1038/520429a

[5] https://sfdora.org/read/

example, we would recommend not using it on groups of less than perhaps 500 papers. Otherwise, adding or removing a single paper can have dramatic effects – even at 500 papers, a single extra "top 1%" paper will shift the overall results by 0.2, quite a substantial change.

As well as the number of papers in your sample, it is useful to consider the size of the overall set of papers you are comparing to. Most normalised indicators such as CNCI and percentile take account of document type as well as subject area, so that there is one baseline level of citations for articles, one for reviews, one for book chapters, etc. Some document types are rare – for example, many fields will not have many book chapters or conference papers – and so may be more prone to being skewed by outliers. In addition, there are some document types, such as book reviews, which are very rarely cited, so the majority have zero citations and a paper with one citation can dramatically outperforms the others.[6] If these documents are included in an overall institution-wide average, they can potentially skew the results.

The best approach to dealing with this is to apply filters for document type and consider which document types are most appropriate for the fields you are looking at.

[6] At the time of writing, November 2019, a 2017 book review in History with only one citation will nonetheless have a CNCI of around 230 and be in the top 0.5% of citations for that document type. An author who happened to write one of these would thus seem to have an exceptionally high average CNCI or average percentile compared to their colleagues, despite a total difference of one citation.

The Indicators

Web of Science Documents

Definition

The number of documents found in the Core Collection[7] matching the selected filters, e.g. publications in a time period.

Key strengths	Key weaknesses
An objective proxy for research production.	Despite large differences in numbers of publications between disciplines, this figure is not field- normalised.
A useful tool to identify rates of research publication between peers in the same field.	Collaboration and multi-authorship make it difficult to assess the contribution of an institution, group or individual to a team effort.

Interpretation

InCites uses whole publication counts, so an output jointly authored by researchers at two different institutions would be credited to both institutions with a value of 1 each. This will often cause multiple counting when comparing groups of outputs. The effect is most pronounced in highly collaborative scientific fields and may be less of an issue in the humanities, where single authorship is more common. InCites does not currently support more sophisticated indexes that take multi-authorship into account using fractional methods.[8] However, there is a benefit to counting citations in whole numbers rather than fractions because this is more representative of how citations function in reality- a citation exists or it doesn't, in practice authors don't receive a quarter or a fifth of a citation.

Practical tips

Using the "Thresholds" filter, it is possible to eliminate any papers with more than a selected number of authors. The appropriate number may vary by field; be aware that massively authored papers are most common in some areas of physics and in article types such as longitudinal studies, but relatively rare elsewhere. See notes on page 27.

The Core Collection covers a broad range of publication types, including different types of articles found within journals. This is one reason why it is important to be familiar with research publication practice in the field you are reporting on. We recommend that you set an appropriate filter for publication types based on the field. The relative status of conference papers, books, and book chapters will vary by field, and so must be decided on a case by case basis- for example up to 40% of an institution's output in History can be book reviews, so the usual advice to exclude book reviews would, in this case, not be suitable.

[7] The Core Collection is a single dataset that exists in InCites regardless of the customer's Web of Science database subscription. For an explanation of the databases that comprise the Core Collection see pages 25-26.
[8] A range of different approaches to weight papers by the share of contribution have been proposed; see Roberto Todeschini and Alberto Baccini, *Handbook of Bibliometrics Indicators*. Weinheim: Wiley Verlag, 2017: 54-74 for a detailed discussion of alternatives to whole publication counts.

You should always clearly state the document type filter if using one. If filtering by number of authors, again, make sure to state the filter used, and why the threshold was selected.

Related indicators

InCites has an indicator for "**Documents in JIF Journals**". This only counts publications in a journal that was indexed in the Journal Citation Reports and given a Journal Impact Factor (JIF) in that year. InCites does not count any papers for the current year as being in JIF journals, as the values are not yet available, and also does not include data before 1996.

In practice, this indicator is somewhat confusing, and we do not recommend using it. In some cases, it can return exceptionally low values; for example, a number of humanities subject categories have zero "documents in JIF journals" (see pages 25-26 for an explanation of the different citation indices), and as noted above it will return zero values for the most recent papers or for those pre-1996.

Lastly, this indicator is discouraged is because its underlying concept is assessing individual publications on the basis of their journal's publication metrics, which is against the principles of the Leiden Manifesto, DORA and many institutional responsible metric policies.

There is also an indicator for "**Documents in Q1-Q4 Journals**" and "**Percentage of Documents in Q1-Q4 Journals**". These indicators show the number (or percentage share) of documents which are in a group of journals ranked by their JIF in a specific discipline category. Q1 journals represent the top quartile (25%) of journals in that field, Q2 journals those in the second quartile, and so on.

Again, we would not recommend using it; the use of journal-based indicators is generally to be discouraged. However, quartiles are a marginal improvement over simply quoting an impact factor; Quartiles have the merits of being broad, do not convey false precision like a formal ranking would, and (as they are defined within Journal Impact Factor categories) are normalised by field.

If using this indicator, it should always be emphasised that journal categories are somewhat arbitrary assignments of discipline to a journal. They may not accurately represent the subject field of every document, particularly multi-disciplinary or emergent discipline work, in the journal. Journals can be assigned to more than one Journal Impact Factor subject group; in this case, they are shown with their top quartile. Where the number of documents in Journal Impact Factor journals is zero, the number in Q1-Q4 journals will also be zero.

Times Cited

Definition

The total number of citations received by a set of outputs since publication.

Key strengths	Key weaknesses
Citations counts are the prevalent indicator of article influence of this age and their essential concept are easily understood by many audiences.	Strong bias towards journals and conferences as primary sources of citations, so this is inappropriate for fields where scholars mainly communicate through other channels.
Citation counts are the basis of understanding the way that citations are made and received across time, fields and geographies and can offer insights into the structure of research and research literature.	Citation rates and lifespans vary wildly in different fields and this indicator is not field-normalised. Never use as the sole indicator when comparing publications in different fields.

Interpretation

The justification for using citations as a proxy for quality or excellence is highly debated.[9] The extent to which citation counts are a credible measure of impact is best judged by experts who understand citation norms and behaviours within the field.

Practical tips

This data covers the time from the paper's first publication to the most recent update of InCites (at time of writing, monthly). It is not normalised in any way by age, and older papers tend to perform much better than newer ones for obvious reasons – they have simply had more time to accrue citations.

Bear in mind that as with all citation data, this figure will include any self-citations and, potentially, negative citations.

We recommend that you only compare the number of citations within the same publication year, the same publication type (articles, reviews, etc) and the same field. Be careful when making comparisons based on small samples, without first flagging extreme statistical outliers.

When reporting total citation counts, make sure to include the date of the most recent InCites update and/or the date of which you collected the data, alongside filters applied to the source data (e.g. "all chemical engineering papers from 2016 to date, counted as of May 2019 and accurate to the April 2019 dataset update of InCites"). Dates of the dataset update can be found at the bottom of any analysis page.

[9]The validity of citations in research assessment is much debated. For an introduction to the topic, see Michael H. MacRoberts and Barbara R. MacRoberts “Problems of Citation Analysis: A Critical Review” *Journal of the American Society for Information Science*. John Wiley & Sons, 40(5):342-349, 1989

Citation Impact

Definition

The number of citations received by a set of outputs divided by the number of those outputs.

Key strengths	Key weaknesses
A size-independent indicator that reveals citation impact relative to the number of outputs.	Caution needs to be exercised to compare like-with-like on the basis of outputs of similar document type with similar age and subject characteristics.
Facilitates benchmarking of broadly similar entities.	Outliers in the data will heavily skew the result because it is calculated by mean.

Interpretation

Outliers in the citation data, such as extremely high or low cited papers, will 'artificially' inflate or decrease the mean. InCites does not provide confidence intervals (a suggested range between which conclusions can likely be trusted) for this indicator but, as a rule of thumb, it is probably sensible to avoid drawing firm conclusions on samples of fewer than 500 outputs.

Practical tips

This data covers the time from the paper's first publication to the most recent update of InCites. It is not normalised in any way by age. This indicator is biased in favour of older publications- simply because they have had more time to accrue citations. We recommend that you only compare the citation impact within the same publication year and the same field.

If you are concerned solely with one type of publication, say articles, then consider using the document type filter to include only the concerned document type. Including all document types will affect the overall average because different types of publications tend to demonstrate different citation averages.

Category Normalised Citation Impact (CNCI)

Definition

The ratio of citations received by a group of outputs against the average for other outputs of the same age, publication type, and subject area(s).

Key strengths	Key weaknesses
CNCI corrects for differences in citation rates attributable to age, document type and subject differences in citing behaviour.	Subject classification is based on a journal-level subject schema and may be inaccurate for some papers.
CNCI effectively benchmarks an output against its immediate peers in the world literature.	CNCI is poorly designed because it uses the mean to characterise heavily skewed citation distributions.

Interpretation

CNCI is a useful indicator for benchmarking purposes because it anchors an output's performance relative to its peers in the scientific literature. An output with a CNCI value of 1 can be interpreted as having 'world average impact'. Like many other citation indicators, the distribution of citations is highly skewed, with very high numbers of citations to a minority of papers. As such, CNCI is best used to benchmark very large publication sets, such as whole institutions, where the effects of statistical outliers are dampened.

The CNCI for a paper will fluctuate over time and is recalculated with each InCites update. However, it is possible to see historic or 'locked' CNCI by selecting the 5-Year Trend Graph as the visualisation and exporting the data in a CSV file. In this file you would be able to see only citation activity within the 5 year time window.

Any paper covered in two or more categories has its CNCI calculated for each, and then the harmonic mean is used.[10] This dampens the effect of very highly-cited papers compared to a simple arithmetical mean.

Practical tips

Data for the most recent two years is potentially quite volatile, and be careful drawing any conclusions from it – the number of citations are often so low that even a matter of months can skew the results. As a rule of thumb, data is most reliable 24 months after publication.

The CNCI attempts to normalise by publication type (article, review, chapter, etc) but this can introduce additional biases. Because many output types are not generally heavily cited, any which do get substantial citations may get incredibly high CNCIs. For smaller groups of papers, this can have a disproportionate effect. It may be worth testing to see the effects of filtering to only including articles/reviews/other highly cited document type.

[10] The harmonic mean is a form of averaging that tends to dampen the effect of very high values (compared with the arithmetic mean.

CNCI is calculated on the basis of the defined **schema**. By default, this is the Web of Science category groupings. If you use another schema, the CNCI will be recalculated on the basis of the new categorisation scheme.

The CNCI can usefully be used to identify very highly cited outliers in a set; clicking on the number of records will bring up the document list view, with the CNCI for each record. You can see how this is calculated by comparing the CNCI value with "category expected citations"; note how this column changes with different research areas and item types.

Good Practice

Field normalisation should always be accepted with a margin of error. It should be sense checked at the concept level for a meaningful comparison.

Comparisons should be made on the basis of similar research mission and scale of output and resources. An appropriate attempt at benchmarking would be, for example, King's College London with Imperial College London outputs in Biochemistry, Genetics & Molecular Biology. An example of inappropriate benchmarking would be comparing all outputs of Imperial College London with all outputs of London School of Economics. The volume and field of outputs is different enough to make the results of comparison meaningless.

As always, make sure to clearly state any filters you are using, and also state which schema you are using for calculating CNCI. Because of substantial differences between the scope of subject areas, the CNCI for an individual paper can vary dramatically when calculated in different schemas.

Related indicators

InCites also offers "**Journal Normalised Citation Impact**" (JNCI), which is calculated in the same way to the CNCI, but on the basis of comparison to other items in the same journal over the same timeframe, rather than in the same field. This avoids the complications of papers being assigned to multiple fields. It gives an indicator of whether papers are "outperforming" or "underperforming" comparable ones from the same publication; a JNCI of 1.25 would indicate that the sample papers are cited ~25% more than others in the same journal.

We would not generally recommend JNCI as an indicator. It only compares papers to those in the same journal – this means that a good paper will score much more highly if it is surrounded by weak papers rather than other good papers. Bearing in mind the prevalent advice against using journal-based metrics for assessing individual publications, it might be better to use a more appropriate metric in the first instance.

Percentage of Documents in Top 1% / Top 10%

Definition

Counts the percentage of a set of documents which are in the top 1% or top 10% of comparable papers (same field, same item type, same publication year)

Key strengths	Key weaknesses
A highly intuitive indicator that takes account of a lot of contextual information (age, document type and field).	Percentile values for recent papers or those in very small groups may be seriously misleading.
Focuses attention at the top end of the citation distribution ("most highly cited papers"). This may be desirable if attempting to demonstrate impact.	Masks the importance of exceptionally highly cited papers (there is still a lot of range within the top 10%).

Interpretation

This is a highly intuitive and easy to understand indicator. All items above the threshold are counted, but only counted once. It thus both emphasises the importance of the most highly cited papers, while muting the effects of outliers with very high numbers of citations.

This indicator complements CNCI and in most cases, it is a good idea to quote both together.

Practical tips

As with CNCI, data from the most recent year or two is highly volatile, and we would not recommend placing much weight on it until it is two calendar years old. The numbers will also be very volatile in topics or institutions with small numbers of papers. We do not recommend using this on groups of less than perhaps 500 papers, especially for "top 1%". Otherwise, the addition or removal of a single paper can cause dramatic changes in the resulting values. Be cautious about what types of outputs are considered. As with CNCI, the use of less common publication types can have a disproportionate effect, and it may become meaningless when article types other than original research and reviews are included.

The indicator only identifies papers in the top 1% and 10%; this can make it tricky to benchmark against analyses which use different percentile groups (e.g. top 5%, top 20%). However, it is possible to carry out this type of analysis using the "percentile in subject area" indicator by downloading the per-paper data and counting manually.

The normalisation will give best results when comparing broadly similar fields and institutions. The values will change when category schema is changed. This is because within each schema the Top 1% or Top 10% count will recalculate to the citation thresholds corresponding to the new category. If a paper is categorised into two subject areas, it is given the percentile of the category in which it scores best. However, it is possible to view both. When opening a dataset in the Research module, select a category and click on the hyperlinked number of records and get to the article level list. For each paper, you will then get the Percentile in Subject Area for the particular category you drilled down, even if the paper is assigned to more than one category. (The logic here is that the user opens

the Research module to investigate the percentiles (and other article level indicators) in each category they like.)

Potentially, and somewhat confusingly, this might mean that some subject areas have *more* than 10% of their reported items in the top 10% grouping.

Related indicators

If a group of documents in InCites is selected to show the 'Web of Science Documents' list view, the indicator "**percentile in subject area**" is displayed in the second to right column This shows the paper's rank among its counterparts (same year, type, and field) in terms of its percentile ranking. The smaller the percentile, the higher a paper's relative citations; a percentile of 10 or less is equivalent to "in the top 10%". This is a reverse percentile. Very highly cited papers may have a percentile value of 0.1 or less. The percentile is calculated on a "best result" basis where a paper is grouped into several different fields.

InCites also offers "**Impact Relative to World**" as an indicator. This compares the overall citation impact of a paper to all papers published globally in that year; the world average is set to 1, so a value of 1.2 would indicate the sample outperforms the global average by 20%. However, this is not normalised by field or document type.

Average percentile

Definition

Gives the average percentile rank of individual papers compared to those in the same subject area.

Key strengths	Key weaknesses
Gives a sense of the overall ranking of a group of papers.	Percentile values for recent papers or those in very small fields may be seriously misleading.
	Does not necessarily indicate actual number of citations yet is a citation-based indicator.

Interpretation

The "**average percentile**" indicator is an aggregation of the "**percentile in subject area**" data, which is calculated for each individual paper by ranking it against all other papers of equivalent age, type, and field. Where a paper is counted in several different categories, its percentile is counted individually for each subject area, and then the strongest value is selected.

The average is calculated as the harmonic mean[11] of all the percentiles of individual papers in the set.

Practical tips

The lower the value, the more highly ranked the paper – this is a "reverse percentile". A value of 50 would indicate the group is exactly average; any value less than 50 suggests the papers are consistently performing above average.

The way this average is calculated tends to dilute the effect of outliers, making it useful to see the overall impact of a set of papers. Similarly, two papers can perform similarly in **average percentile** yet have very different actual citation counts.

Displaying this alongside the top 1%/10% and CNCI figures can give a rounded picture of the overall citation patterns.

Related indicators

If a group of documents in InCites is selected to show the 'Web of Science Documents' list view, the indicator "**percentile in subject area**" is displayed in the second to right column. This shows the paper's rank among its counterparts (same year, type, and field) in terms of its percentile ranking. The smaller the percentile, the higher a paper's relative citations; a percentile of 10 or less is equivalent to "in the top 10%". This is a reverse percentile. Very highly cited papers may have a

[11] The harmonic mean is a form of averaging that tends to dampen the effect of very high values (compared with the arithmetic mean.

percentile value of 0.1 or less. The percentile is calculated on a "best result" basis where a paper is grouped into several different fields.

It is possible to see the other field percentiles and not just the ‘best result”. When opening a dataset in the Research module, select a category and click on the hyperlinked number of records and get to the article level list. For each paper, you will then get the Percentile in Subject Area for the particular category you drilled down, even if the paper is assigned to more than one category. (The logic here is that the user opens the Research module to investigate the percentiles (and other article level indicators) in each category they like.)

% Highly Cited Papers and Highly Cited Papers

Definition

This indicator shows either the percentage or absolute number of papers that are classified as "highly cited" by the Essential Science Indicators (ESI) database (a separate paid-for database service from the owners of InCites).

It can be easy to confuse this indicator with the similar, but not identical standard InCites "Top 1%" indicator (explained on page 12)- try to avoid confusing the two!

Key strengths	Key weaknesses
ESI is a selective and well-regarded dataset	ESI groupings are quite broad – only 22 subject categories
Intuitive indicator that focuses attention at the top end of the citation distribution ("most highly cited papers"). This may be desirable to show impact.	Not all papers are classified in the ESI data – it is a smaller corpus of data than InCites and built using a different methodology[12].

Interpretation

This is a reasonably intuitive indicator. It shows the number (or proportion) of papers classified as 'highly cited' by the Essential Science Indicators database that the selected entity has produced in a given time period. The classification of 'highly cited' in the ESI is defined as the top 1% of cited papers, per year based on the previous ten years, within each of the 22 subject areas comprising the ESI. It is calculated bimonthly (six times a year).

As with top citation percentiles, it counts all papers above the citation threshold, but only counts them once.

The ESI dataset is smaller and more focused than the standard InCites dataset; it includes only articles and reviews from journals in the SCIE and SSCI, not those in the AHCI or ESCI.[13]

Practical tips

Highly Cited Papers are only calculated for the last ten years; after this, no papers are categorised as highly cited and values will always be zero. This can cause odd results if averaged over a long period of time.

Bear in mind that the categorisation used by ESI is very broad, just 22 categories as opposed to the hundreds that may be available in other schemas. Particularly for smaller cross-disciplinary subjects, they may be misleading compared to using the citation percentiles with an appropriate schema. In such cases, we would recommend using InCites' "% in Top 1%" option instead.

[12] For an explanation of the construction of the ESI see page 26.
[13] ESI coverage is outlined at https://help.incites.clarivate.com/incitesLiveESI/ESIGroup/overviewESI/scopeCoverageESI.html

ESI only considers journal articles and reviews, not conference proceedings or other formats; if the field in question makes heavy use of these (e.g computer science) then the results may be misleading.

Be careful when selecting whether to use the number of Highly Cited Papers, or the percentage. The percentage is a more meaningful result (as it is effectively normalised by the number of papers produced by an entity) but, as with citation percentiles, it can be very volatile for smaller groups of papers. If using the absolute number, it is wise to qualify it (eg "25 out of 2,000").

Related indicators

A similar indicator is **% Hot Papers**[14], which is also calculated using ESI data.

For organisations only, the indicator "**ESI Most Cited**" is available. This calculates all papers assigned to the organisation for the previous 10 year period, normalised for each of the ESI 22 research areas, and indicates 'YES' if the organisation has produced a paper that is classified in the top 1% of any research area (and 'NO' if it has not). As demonstrated in Figure 1 below all organisations that are indicated 'YES' are ranked 1 (and all indicated 'NO' are ranked 0). The results are displayed by country (United States is the default) then within the large ranks of 1 and 0 the ordering seems to be somewhat alphabetical by organisation name. The use of this indicator is strongly discouraged- it is difficult to see what useful meaning could be derived from this.

Name	Rank	Web of Science Documents	▼ ESI Most Cited
University of Alabama Tuscaloosa	1	29,120	YES
University of Alabama Huntsville	1	11,396	YES
University of South Alabama	1	15,958	YES
University of Alaska Fairbanks	1	17,396	YES
Arizona State University	1	97,349	YES

Figure 1 Screenshot of Organisations ranked by ESI Most Cited indicator

[14] See pages 18-19 for analysis of this indicator

% Hot Papers

Definition

Shows the percentage of papers that are classified as "hot papers" by the Essential Science Indicators (ESI) database (a separate paid-for database service from the owners of InCites). Hot papers are the top 0.1% of papers published in past two years and cited in the past two months.

Key strengths	Key weaknesses
Within the subject categories defined it can identify highly cited papers and emerging research trends.	ESI subject categories are broad – only 22 subject categories which is not representative of the diversity of research themes.
Intuitive indicator that focuses attention at the top end of the citation distribution ("most highly cited papers in recent months"). This shows emerging fields of research.	Limited to articles and reviews from the SCIE and SSCI.

Interpretation

This indicator is conceptually similar to the other ESI indicators, but is very narrowly defined. It identifies papers which were published in the last two years, and counts those which have been most heavily cited in the last two months – the top 0.1% of papers. The narrow definition means that a very large sample set would be needed for meaningful results. As with top citation percentiles, it counts all papers above the minimum citation threshold, but only counts them once regardless of how heavily they were cited.

ESI data is updated bimonthly (six times a year) and this will be reflected in monthly InCites updates. Because of this, % Hot Papers is an extremely volatile indicator.

As with the ESI Highly Cited Papers indicator, this only considers articles and reviews in science and social-science journals.

Practical tips

This is a very time-limited indicator, it will only give meaningful results for recently published papers.

Bear in mind that the categorisation used by ESI is very broad, just 22 categories as opposed to the hundreds that may be available in other schemas. Particularly for smaller cross-disciplinary subjects, they may be misleading compared to using the citation percentiles with an appropriate schema. In such cases, we would recommend using InCites' "% in Top 1%" option instead.

ESI only considers journal articles and reviews, not conference proceedings or other formats; for any field that makes heavy use of these (e.g. computer science) the results may be misleading.

At the time of writing, the Hot Papers indicator is only available as a % of Hot Papers, no raw number. If the raw number is desired this could be found by searching in the Web of Science or the ESI databases. The raw number if presented should be contextualised with further details (e.g. “25 out of 2,000 papers”)

Related indicators

Hot Papers is conceptually similar to the **Highly Cited Papers** and **Percent Highly Cited Papers** indicators[15]; these also use the ESI data but use a different threshold (top 1% of papers) and all citations, not just those in the past two months.

When looking at organisations, there is a filter option for "**ESI Most Cited**"[16].

[15] See page 16 for analysis of this indicator
[16] See page 14 for analysis of this indicator

% of Documents Cited

Definition

This is a simple indicator that reports the proportion of documents which have one or more recorded citation.

Key strengths	Key weaknesses
A very simple and straightforward indicator	In many fields, having one citation is a very low threshold. The data may thus not be very meaningful.
	Older papers are naturally more likely to have been cited than newer ones, but this indicator is not normalised by age.

Interpretation

This indicator shows the proportion of publications which have received at least one citation. It is one of the default indicators shown by InCites, and so may be quite commonly encountered.

Practical tips

This indicator is not normalised by age, by field, or by publication type. All of these factors can affect the results – most obviously, older papers are naturally more likely to have been cited than newer ones, so percentages will tend to be quite high for older publications. In addition, the mean time-to-first-citation varies dramatically between fields; in some fields only 20% of papers are cited in their first two years, while in others 90% will have been cited by this point. Finally, different types of publications will tend to have different citation patterns; some, like book reviews, are almost never cited, but make up a substantial proportion of the outputs recorded by InCites in fields such as history. Because this figure is expressed as a proportion of the total, including rarely-cited material will tend to drive down the average.

You should take care to only compare similar fields, groups of papers of similar age, and set the article type filter carefully. It is not possible to normalise for these factors, and the only option to normalise this data is to run individual queries for small groups of papers – for example, all papers in 2016, all papers in 2017, etc.

If choosing to use this indicator after following the advice above to 'self-normalise' then one should also consider the concept of this indicator, especially at the lower or higher ends of citation count. There is essentially little difference between a paper with one citation and a paper with zero, or between a paper with 1,000 citations and 1,001 citations in terms of assessing significance or impact.

Additionally, studies have suggested that a significant of "uncited" papers have actually been cited but the citation indexer (in this case the Web of Science) hasn't identified the citation.[17]

[17] Van Noorden, R. The Science that's never been cited. *Nature* (2017) 162-164 doi:10.1038/d41586-017-08404-0

International Collaborations

Definition

The number of documents that have one or more co-author affiliated to another country (one that is different to the country of the person, organisation, funding agency, region or research area that you have selected).

Key strengths	Key weaknesses
Simple to understand	Some anomalies in country assignation data, and dual affiliations can be identified as "collaboration"
	Boolean indicator- a single document either *does* or *doesn't* have a collaborator assigned to a country different to that of the entity you have selected or not. It is not an indicator of the number of or diversity of collaborators beyond this.

Interpretation

This indicator shows the number of papers which have internationally collaborating authors. Papers are counted this was if they have any two listed affiliations from different countries, regardless of the number or share of authorship – a paper with sixteen authors from sixteen countries will count the same as a paper with fifteen from one country and a single author from a second. An author with dual affiliations in different countries will also make a paper count as internationally collaborative.

Practical tips

InCites does not have perfect affiliation data, e.g. authors can have multiple affiliations, complex co-author networks, change between affiliations or be affiliated to organisations that are not recognised by this dataset. Whilst InCites does attempt to unify and disambiguate authors, this is notably a mammoth task that many indexers and analytics products struggle with.

A small percentage of papers do not have any country assigned to authors and so will not be covered here.

Consider the disciplines covered in your publication set. Documents from a field that is typically collaborative, such as high energy physics, will be more likely to perform well in this indicator as opposed to a field that produces a lot of single author works, such as history. Applying this indicator to a publication set that includes different fields may thus be misleading.

It is possible to apply a threshold on the number of authors to exclude or focus on large multi-authored papers, this could facilitate the study of single author or collaborative works in certain fields. The threshold available 'Authors per Document' refers to the number of authors not the number of countries but this could still be a useful filter to combine with this indicator for more meaningful results. What constitutes a "normal" level of collaboration as opposed to an unusually high level will vary by field, so bear this in mind when selecting the threshold if it covers a range of disciplines.

If using the Person interface, the 'home' country affiliation of the author will be counted in the results.

Related indicators

This indicator can also be expressed as a percentage - "**% of international collaborations**", e.g. the percentage of papers that count as internationally collaborative.

A related indicator is "**% of industry collaborations**"; rather than showing the share of papers with an international co-author, this shows the share with an author from an identified corporate affiliation. This should be used with extreme caution. InCites note that that their non-academic organisational data is still in development; it is likely to undercount the number of corporate affiliations and may be skewed towards identifying large multinationals over smaller or regional enterprises.[18]

[18] *InCites* Indicators Handbook, p.19

h-index

Definition

The h-index of the set of papers being reported on, where h is h papers in the set have at least h citations reported.

Key strengths	Key weaknesses
The h-index is a well-known metric and the calculation behind it is easy to understand and reproduce.	There is wide misuse and a lack of understanding of the necessary caveats for practical use to learn anything meaningful from this indicator. For a brief introduction to the limitations of this indicator see The Metric Toolkit[19].

Interpretation

Shows the h-index of the set of papers being displayed by InCites. It is most commonly used for individuals but can also be calculated for any selected group of papers.

Practical tips

The h-index should not be used for comparing institutions, funders, or journals. These tend to have very different numbers of papers, and all other things being equal, the h-index is directly related to publication volume.

When used for individuals, it tends to correlate closely to the age of the researcher, the length of career, including career breaks and can be heavily influenced by historic publishing patterns.[20] It is also heavily dependent on the field, as some fields tend to produce more papers and citations than others, thus making even sub-discipline level comparison invalid. As an indicator of people, the h-index is of limited value and should be treated with great caution.

In addition to the general problems, InCites itself has incomplete attribution of papers to authors, with some people incorrectly grouped together as one and others split apart.[21] This makes calculations based on individual publication records quite unreliable; one missing paper or even one missing citation can have a substantial effect on the outcome for a small sample set. Unless you are very confident about the accuracy of the data, it may be safest to avoid using it at all.

[19] https://www.metrics-toolkit.org/h-index/

[20] There is a large amount written on the flaws of the h-index. See, for example, a discussion by Ian Rowlands at https://thebibliomagician.wordpress.com/2018/03/23/is-it-time-to-bury-the-h-index/

[21] Problems identified in correct attribution of authors include commonly-shared names, different naming order conventions, transliteration of names between alphabets and languages, changes of author names and more.

InCites data sources

Bibliometric Data

InCites data comes from the Web of Science Core Collection. This is a selective database of sources including journals, conference proceedings and books. The entire Web of Science Core Collection is represented in InCites, regardless of what individual elements are in the institutional Web of Science subscription. However, one part (the Emerging Sources Citation Index) is optional and can be toggled on and off as preferred and all citation counts will include only citations that have come from journals indexed by the Core Collection.

All of the sources present have passed a quality evaluation by a Web of Science editorial panel and can be removed from the database ('deselected') by the editors if they perceive a failure in editorial or scientific standards, although this is rare and only a small number of journals are withdrawn each year. As a highly selective database, the Core Collection records only titles that have passed their evaluation process. This should be made explicit when presenting InCites data to a non-expert audience, especially if they are familiar with more inclusive and generally larger datasets like those that power Scopus or Google Scholar.

Studies of the Web of Science have found that the admission criteria for subject coverage and geographical/language coverage skew the journals which are included in the database. Arts and humanities research is underrepresented in the Core Collection.[22] Studies comparing the selective evaluation for Web of Science admission with a non-selective database, Ulrichsweb, also suggest that Web of Science coverage underrepresents journals that do not publish predominantly in the English-language.[23] The concentration on journals from major commercial publishers (headquartered in the United States, the United Kingdom and the Netherlands) also results in substantial overrepresentation of journals published in those regions.[24]

Coverage of material outside journal articles is more limited, with less comprehensive coverage for conference papers, books and book chapters, and no coverage of working papers or preprints. This has implications for fields which rely more heavily on these as publication venues; not only are fewer publications included in InCites, but their citation counts are similarly affected as fewer citing documents are recorded.

Institutional Data

InCites states that there are three sources of institutional data in the product. Guidance on understanding the origins and implications of the institutional data is below.

Declared data source	Interpretation
Institutional data collection	InCites state that they collect data directly from institutions[25] as well as 'Where appropriate, we utilize authoritative third-party data sources and all of the data is validated by the institution'. The data is suggested to be collected "once per year… We invite roughly 1,200 organisations; typically

[22] Mongeon, P. & Paul-Hus, A. Scientometrics (2016) 106: 213. https://doi.org/10.1007/s11192-015-1765-5
[23] Archambault, É., Vignola-Gagné, É., Côté, G. et al. Scientometrics (2006) 68: 329. https://doi.org/10.1007/s11192-006-0115-z
[24] ibid Mongeon
[25] http://ips.clarivate.com/globalprofilesproject/ Assessed 06.06.2019

	around 700-800 participate"[26]. It is not possible to filter between the data sources in InCites so what the end user sees will be a mix of what is provided by an institution, and where this is missing, data acquired from a publicly available source, for example HESA for UK-based institutions.
Clarivate Analytics Annual Academic Reputation Survey	A survey described by InCites as a poll conducted in April-May annually of 'academics around the world asking their feedback on what they consider to be the strongest universities and research institutions in their discipline' to a closed invitation-only pool of respondents whom 'on average…have more than 16 years experience working in academia'[27]. Whilst Clarivate Analytics release a summary of the findings annually, the most recent information on the methodology that is publicly available dates to 2010[28].
Bibliographic analysis	Derived from InCites publication and authorship data, e.g. calculating the spread of co-author countries to derive a proportion of papers which do or do not have international co-authorship.

The citation indexes

The citation indexes within the Core Collection are databases structured by subject area, 'categories'. Journals are assigned categories when they are admitted to the Core Collection. Note that for some journals their category assignation could have occurred years or evens decades ago when the journal was first admitted (the category assignations are sometimes reviewed at the request of the publisher or independently by the Web of Science editors). This could pose a problem for any indicator normalised by journal field to assess research in multidisciplinary or emerging fields, as they may not be accurately represented by the database of sources and citations selected.

Science Citation Index Expanded (SCIE)

The largest database, it indexes journals that are assigned to categories that generally fall within traditional classifications of the sciences. Multiscience journals will feature here. All journals in the SCIE are awarded a Journal Impact Factor.[29]

Social Science Index (SSCI)

Indexes journals that are assigned to categories that are traditionally acknowledged to be social science. Generally, both applied and theoretical research will be represented in these journals. All journals in the SSCI are awarded a Journal Impact Factor.[30]

Emerging Sources Citation Index (ESCI)

This is a new index in the Core Collection and contains citation data from journals that have been identified by Clarivate Analytics editorial teams as being 'important' titles[31] and have passed the quality criteria set by the editors for admission to the Core Collection.[32] All journals in the ESCI are indexed with full citation records; citations from them will contribute to the Journal Impact Factor calculations for Core Collection journals, but ESCI journals themselves are not awarded Journal

[26] Information provided to authors by Clarivate Analytics representative in email correspondence dated 15.06.2019
[27] Ibid
[28] http://ips.clarivate.com/m/pdfs/GIPP_AcamRep_methodology.pdf Accessed 06.06.2019
[29] It is possible for a journal to not be awarded a Journal Impact Factor - it could be too 'young' (e.g. not having a prior two years' worth of published content) or if in very rare cases the Journal Impact Factor has been suspended
[30] See above
[31] http://info.clarivate.com/ESCI Accessed 06.06.2019
[32] https://clarivate.com/essays/evaluation-criteria-web-science-core-collection-journals/

Impact Factors. The journals indexed by the ESCI are continually evaluated for inclusion in a main index (such as the SCIE or SSCI), at which point they would be awarded a Journal Impact Factor. The ESCI does not have disciplinary classifications.

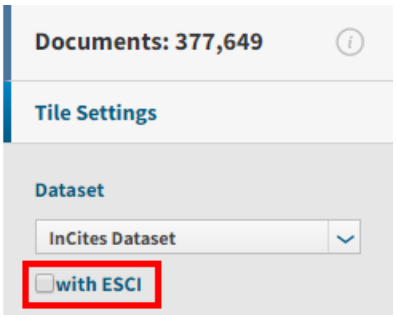

Figure 2 InCites menu bar with 'with ESCI' checkbox highlighted

Users have the option to include ESCI data or to exclude it, whether working with either an InCites Dataset or Web of Science data. The default is to exclude the ESCI data. Users are encouraged to be conscious of the 'with ESCI' tick box that appears in Tile Settings and to check regularly that it is still ticked or unticked as per data preference. We recommend that users include ESCI data in their analysis because it broadens the number of sources.

Other Indices to be aware of

Journals indexed in the Arts & Humanities Index (AHCI) are not awarded Journal Impact Factors but citations from them will contribute to the Journal Impact Factor calculations for Core Collection journals. The Books Citation Index database exists for Science (BKCI-S) and Social Science & Humanities (BKCI-SSH) and will capture citations from scholarly books, which can be substantial in some fields. The Conference Proceedings Citation Index Science (CPCI-S) and Social Science & Humanities (CPCI-SSH) index conference proceedings, which for some fields, computer science for example, can be highly important.

Some indicators (Highly Cited Papers, Hot Papers) use the "Essential Science Indicators" (ESI) dataset. This is based on purely articles and reviews from journals in the SCIE and SSCI; it does not include any other material. The ESI dataset is also limited to the past ten calendar years, so older papers are not included. [33]

Selecting your dataset - InCites Dataset or Web of Science?

A new analysis will open with the default dataset as 'InCites Dataset' and excluding the Emerging Sources Citation Index (ESCI). However, users who wish to analyse a specific set of records that have been identified through the more sophisticated search and filter function of the Web of Science can do so by exporting directly into InCites. Note: there is a limit of 50,000 records in the export and papers published before 1980 will be excluded.

Users also have the option to upload a customised dataset in a CSV/TXT file containing a list of DOIs, Web of Science Accession Numbers and/or PubMedIDs at a rate of up to 50,000 records per upload.

[33] The ESI content is outlined at https://help.incites.clarivate.com/incitesLiveESI/ESIGroup/overviewESI/scopeCoverageESI.html

Data selection matrix

The following chart identifies data implication of each option[34].

	InCites Dataset	Web of Science data
Without ESCI	The full InCites dataset extends back to 1980 and includes content indexed in: • Science Citation Index Expanded • Social Sciences Citation Index • Arts & Humanities Citation Index • Conference Proceedings Citation Index Science & Conference Proceedings Citation Index Social Sciences & Humanities • Book Citation Index Science & Book Citation Index Social Sciences & Humanities	The results of a Web of Science search can be exported directly into InCites. If selected, this would limit your InCites analysis to the results found in your Web of Science search, which themselves are **dependent on your Web of Science subscription.**
With ESCI	As above, with complete Emerging Sources Citation Index 2005 - present	As above, with complete Emerging Sources Citation Index 2005 present (**available only to ESCI backfile subscribers)**

Setting thresholds for your dataset

In addition to the filters which you can set by subject area, institution, etc, InCites offers a set of "thresholds" to limit and exclude. These allow you to limit the results returned.

- **Web of Science Documents** – this will limit results to entities with the specified number of Web of Science publications.
- **Times Cited** – this will only show results with total times cited in the specific range. (This is not the same as the number of citations on a given paper.) Remember that this indicator is not normalised by age or subject area, and is simply an aggregate total.
- **Authors per Document** – this will filter papers on the number of recorded authors.
- **JIF Quartile** – this will filter papers on the impact factor quartile of their journal.

The first two can be useful for setting minimum thresholds in order to ensure a list only contains significant results – those with more than a minimum number of papers or more than a minimal level of citations. We would recommend using them to refine an existing search, rather than applying them before seeing what the initial results are.

The third, number of authors, can be very useful in some circumstances. It can be used to remove "hyper-authored" papers with many authors – these are very common in some fields, such as particle physics, but in others may be anomalies that should be filtered out. This is particularly relevant when analysing collaborations, as hyper-authored papers are likely to identify a lot of "collaborations" that may not reflect actual individual contacts.

The threshold for what constitutes hyper-authorship varies by field and in different contexts; a commonly used value is 50 authors, but in some fields it may be appropriate to set it lower, at perhaps 15 or 20.

[34] Adapted from http://clarivate.libguides.com/incites_ba/aboutdata Assessed 15.05.2019

The fourth, JIF quartiles, is probably best avoided due to the well-known problems of assessing publications by the ranking of their journal. It also suppresses any publications in journals without impact factors, as these do not belong to any quartile.

Research Area: the schemas

Under "Research Area", users can select a schema to group output results based on their subject. The most commonly used schema is the Web of Science category schema; this represents around 250 subject categories,[35] and will be the schema used if you do not change the selected field.

Those used to working with research data will be familiar with some of the specific schemas, for example the UK RAE/REF subject classifications, relevant to recent UK national research quality assessments.

The selected schema is critical for the calculation of subject-normalised indicators like the category-normalised citation impact (CNCI) or the group of papers in the top 10% of publications, as Individual schemas are calculated on the basis of existing Web of Science category groupings. As an example, we will look at the three REF-related schemas – RAE 2008, REF 2014, and REF 2021. These are based on the units of assessment used in the three UK research assessment exercises, though in 2008 and 2014 a small number of units of assessment were not included in the schema.[36]

Each unit is assembled out of a number of different Web of Science categories; for 2021 for example, unit 1 (clinical medicine) contains 31 different categories from allergy through to virology. At the other extreme, unit 15 (archaeology) contains only one category, and so maps exactly to the Web of Science schema for "archaeology". Papers which are assigned to "multidisciplinary sciences" in the Web of Science are included in a special "not classified" category.

A small number of categories are assigned to two units, meaning for example all papers classed as "Biotechnology & Applied Microbiology" are assumed to fit into REF units 5 (biological sciences) and 6 (agriculture, food, & veterinary science).[37]

As discussed above, it is important to stress that these are journal-based classifications that may not directly reflect the subject of the papers due to the level of aggregation involved. In practice, many papers are submitted for assessment under a different unit to the one notionally assigned to their journal, highlighting the discrepancy.

[35] There are 254 categories as of January 2019; details at https://clarivate.com/webofsciencegroup/blog/new-web-of-science-categories-reflect-ever-evolving-research/. The exact number varies slightly over time as new subject categories are created, and existing ones split.

[36] For 2008, this is unit 5 ("other laboratory based clinical subjects") in Panel A, and unit 8 ("primary care and other community based clinical subjects") in Panel B. For 2014, this is unit 15 ("general engineering") in Panel B. All 36 categories in REF 2021 are represented in the InCites schema.

[37] In 2021, 20 categories are assigned to two units (none are assigned to more than two). In 2014, this was 21 categories assigned to two or more units. In 2014, with a much larger number of units overall, there were 74 categories assigned to two or more units, with some assigned to as many as five.

Comparing indicators in InCites and SciVal

There are multiple providers of research data and analytics. A rival product to InCites is Elsevier's SciVal.[38] The purpose of this section is not to compare the two as overall products but to provide guidance at the individual indicator level.

When thinking of where equivalents exist or don't exist between two providers of metrics, the Snowball Metrics initiative[39] can be useful. Where either InCites or SciVal declare they use a Snowball Metrics-compliant methodology for an indicator it will be indicated with an (*) below.

InCites indicator	SciVal equivalent	Comments
# Web of Science Documents *	Scholarly Output*	Both will include journal publications (including various article types) as well as books and book chapters. As of June 2019 InCites update, InCites indexes close to 60.5 million records and SciVal 74 million in the same month. Both data sources come from each's respective primary databases, Web of Science Core Collection and Scopus.[40]
Times Cited*	Citation Count*	
Citation Impact*	Citations per Publication*	See above.
Category Normalised Citation Impact (CNCI)*	Field Weighted Citation Impact (FWCI)*	SciVal state that an output's FWCI is 'locked' after four years. Follow the advice in the CNCI indicator section of this document for details on how to access 'locked' values in InCites
Journal Normalised Citation Impact	n/a	
% of Documents in Top 1% and % of Documents in Top 10%*	Outputs in Top Citation Percentiles*	
Percentile in Subject Area	n/a	Can be considered as near-similar to Outputs in Top Citation Percentiles with 'Show as field-weighted' included as described above.
Impact Relative to World	n/a	
Highly Cited Papers	n/a	
Hot Papers	n/a	
% Percentage of Documents Cited	n/a	

[38] For SciVal indicators please refer to Elsevier's Research Metrics Guidebook (https://www.elsevier.com/research-intelligence/resource-library/research-metrics-guidebook) and the 'Using SciVal Responsibly Guide' companion guide to this, email Robyn.price@imperial.ac.uk for details

[39] Snowball Metrics is a community project to define and agree metrics and methodologies that can be adopted by any institution or data provider and is aimed at creating a global standard of data practices that will lead to better benchmarking and thus better decision-making. https://www.snowballmetrics.com/

[40] For analysis of the journal coverage in each database, see Mongeon, P. & Paul-Hus, A. Scientometrics (2016) 106: 213. https://doi.org/10.1007/s11192-015-1765-5

International Collaborations	Collaboration Impact*	SciVal presents Collaboration Impact in a % and broken down to international, national and single authorship. Then you must select each of these 'levels' of collaboration in Scholarly Output to view the publications by individual country.
% of Industry Collaborations	Academic-Corporate Collaboration*	
h-index	h-index*	

Creating organisational groups (e.g. departments)

One limitation of InCites is that, in most cases, it can only give information on an institutional level. This is a problem for users who only want data relevant to a single group within an organisation, e.g. a department or research centre.

InCites offer an additional paid for module (*MyOrganisation*) which allows an institution to integrate its organisational structure into InCites, and so produce customised reports based on individual faculties or departments. If this tool is not available, however, it is still possible to produce a similar analysis manually, using the InCites custom datasets function.

You can generate papers by using **a custom search in Web of Science**, for example by using a custom affiliation filter to find relevant papers. From the Web of Science search results page, select "save to InCites" from the drop-down menu.

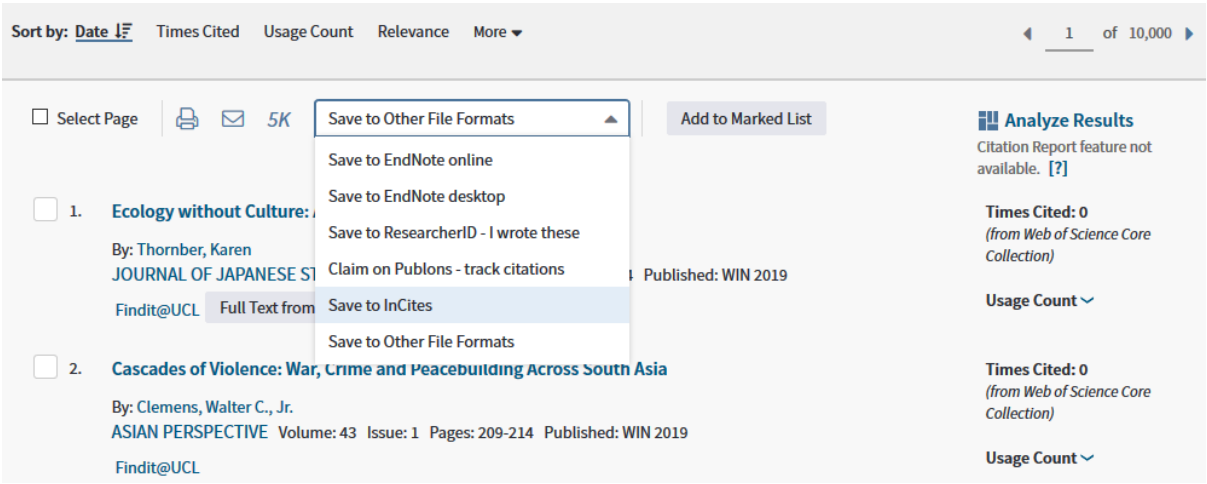

Figure 3 Screenshot of Web of Science 'Save to InCites' function

You may be asked to log into InCites at this point. The new dataset will take a few minutes to copy into InCites, and you will be notified by email when it is ready.

As of 2019, there is a hard limit at this point of 50,000 records – it is not possible to send a larger dataset to InCites. If your dataset is larger than this, consider whether it could practically be subdivided and analysed in parts (perhaps by year). There is also a storage limit of 20 custom datasets – if you have exported too many you will need to delete old ones within InCites before adding more.

It is worth noting that InCites covers almost all papers from the Core Collection but does not have absolutely everything; the update lag means that it omits very recent papers, and some smaller journals are in the optional "emerging sources" index, which can be added back in with the "with ESCI" filter. In addition, InCites only begins at 1980, so older papers will be discarded. As a result, the total number imported into InCites will generally be a little smaller than the reported number in Web of Science – this is not a cause for concern.

After exporting your data, open InCites as normal, and open your preferred analysis option (e.g People, Organizations etc). Then select the new dataset in the left-hand menu and let it load. You may need to reset the filter options after selecting the new dataset – take care to make sure that all your filters are exactly as you want, including whether or not to include ESCI.

You can now generate aggregated statistics for your whole dataset by using the dataset baseline options – these will reflect all papers in the dataset, taking account of any filters. These figures can now be compared to the overall values for the entire institution (obtained through the InCites main dataset), or to those for other custom datasets, as appropriate.

An alternative approach is to use a **pre-generated list of identifiers**, perhaps exported from your local research information software. With your list of identifiers saved, you can click 'My Folders' > 'Create' > 'Dataset' > 'Choose File' and upload the file.

The file should be a plain-text list of identifiers, one on each line. DOIs do not need to be prefixed; Web of Science identifiers should be prefixed with WOS:, and Medline IDs with MEDLINE:. Bear in mind that not all DOIs will be found in Web of Science, and that some may be duplicated – this is most frequently the case with book chapters, which often have a single DOI for the full volume, shown on each chapter record.[41] Duplicate IDs, and items not found in the InCites database, will be ignored.

If you find that your data is returning substantially fewer results than anticipated, you can try loading it into Web of Science to assess what the matching looks like. To do this, construct a Web of Science advanced search like:

```
DO="10.1234/…" OR DO="10.1234/…" OR (etc)
```

where each DOI is entered separately. Use `UT="…"` for Web of Science IDs. The search can cope with a very large number of individual strings – at *least* ten thousand at a time. If you hit this limit, then run several searches and combine them.

As well as trouble-shooting the matching, this approach can also be used to apply filters to a set of papers before migrating them into InCites, should the relevant filters/searching only be available in Web of Science.

[41] Full details on this process are at https://help.incites.clarivate.com/inCites2Live/myFoldersGroup/customDatasets.html

Disclaimer

InCites is a registered product of Clarivate Analytics. The authors of this document do not warrant the accuracy or completeness of the information provided. The intention of the authors is to offer guidance on the use of indicators contained in InCites, their meaning, application and limitations.

About the Authors

Andrew Gray is the Bibliometric Support Officer at University College London, where he has been responsible for the development of an institutional policy on the responsible use of metrics, and supporting researchers with bibliometric enquiries. He has worked as an academic librarian specialising in scholarly communications since 2013. Contact at andrew.gray@ucl.ac.uk

Robyn Price is the Bibliometrics and Indicators Manager at Imperial College London, where she is responsible for establishing and managing a bibliometrics service with a focus on promoting responsible use of metrics. As a member of the LIS-Bibliometrics Committee she manages The Bibliomagician. Robyn's previous experience is in academic publishing, having worked in the editorial teams of subscription and open access journals. Contact at robyn.price@imperial.ac.uk.

Further Support inc LIS-Bibliometrics

A community of support for bibliometric practitioners can be found on the LIS Bibliometrics JISC Mail discussion list (https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=LIS-BIBLIOMETRICS) and The Bibliomagician blog (https://thebibliomagician.wordpress.com/).

Version details

Version 1 published January 2020.

Available at https://doi.org/10.25561/75946.

License info