Lesson 2.4: Cleaning up publisher name variants

# Required reading

* Vincent Larivière, Stefanie Haustein, and Philippe Mongeon, "The Oligopoly of Academic Publishers in the Digital Era," *PLOS One* (2015), <https://doi.org/10.1371/journal.pone.0127502>: This article describes the consolidation of the publishing market over the last 50 years and finds that the five largest publishers accounted for more than 50% of all papers published in 2013. The changes in the publishing market described by this article account for many of the publisher name variants you’ll clean up in this lesson.

# Recommended reading

* [NISO Journal Transfer Alerting Service](https://journaltransfer.issn.org/): The Transfer Alerting Service exists to share information about transfers of journals from one publisher to another.

# Objective

* Create a list of standardized publisher names based on the list of publisher name variants from your dataset.
* Map the standardized names into your dataset and analyze your dataset.

# Key points

* The dataset you built in Section 1 has a publisher listed for each item. Publishers can have multiple imprints, and individual data sources often use several name variants for a single publisher. Publisher data needs to be "cleaned" to develop an accurate set of data for publications from a single publisher.
* The need for this cleaning increases with the number of items in your dataset—the more publications, the more chance for a new publisher name variant to arise. If you have a relatively small number of publications in your dataset, you may not need to follow all of the steps here. Read through the lesson and see what will make the most sense for your situation.
* Different data sources will list different publishers for particular journal titles. A journal’s publisher can also vary over time. See the section "Sources of inconsistency and inaccuracy in publisher data," below.

# Lesson Outline

[Background](#_196m5s9h21k6)

[When you need publisher data](#_khdmpctv71sg)

[What you need to clean up](#_a3p3yfqfgvf)

[Sources of inconsistency and inaccuracy in publisher data](#_bw9xxjbz8q5p)

[Procedure](#_vnpmj3trgt5k)

[OpenRefine](#_jliupc6ulu5)

[Making your own list of standardized publisher names](#_ul8tdfmybe6u)

[Text filter on publisher name](#_ijszw6rx28i8)

[Match to Unpaywall publisher field](#_d66w4q268j0e)

[Clustering](#_hmre533x00p)

[Mapping standardized publisher names into your dataset](#_aeq6odeqr1bc)

[Analyzing your data](#_5s0k0xbu0lwm)

[Excel](#_257aetxottjp)

[Making your own list of standardized publisher names](#_ih0jqxdnh7kd)

[Mapping standardized publisher names into your dataset](#_ofaje84t2rgp)

[Analyzing your data](#_j06js1y7vz42)

[Appendix: Publisher imprints and other names to standardize](#_boz3mp7eya1c)

[Elsevier](#_ml0uai42ulw2)

[Springer Nature](#_bk1lowijylf7)

[Taylor & Francis](#_e07xlmn3wrme)

[Wiley](#_xgm3ug3vs8dj)

[De Gruyter](#_byzqwtor1lzo)

[Wolters Kluwer](#_hdtcx6mvbdmj)

[Other publishers](#_5xpy4lwqz7bu)

# Background

## When you need publisher data

Depending on your goals, you may not need to work with publisher data at all. For instance, having detailed publisher data is not necessary if you’re reporting on all of an institution’s output in aggregate. However, knowing which articles come from a given publisher can be helpful.

For example, you may wish to contact publishers with whom authors at your institution frequently publish, perhaps to inform them about a new OA policy or begin an OA-related negotiation. Identifying the largest publishers of your institution’s research can help you plan outreach to or negotiation with publishers.

Or you may be considering whether to renew journals from a particular publisher and want to know whether researchers from your institution publish in them. Keep in mind that journals not indexed in your data source won’t show up in your data but may publish articles from your institution’s authors. You may also miss titles that have recently been transferred (although if you have a list of these titles, often called takeover titles, you can search for them individually).

## What you need to clean up

If you’re working with the publisher data in your dataset, you need to clean up publisher name variants. Many data sources use more than one term to refer to a single publisher, either due to mistakes or because they refer to imprints.

Imprints are units within a larger publisher, operating under trade names. For example, Academic Press is an Elsevier imprint. It is unlikely that your library has a relationship with Academic Press specifically. Instead, you would probably subscribe to that content, if you wanted it, through Elsevier. Similarly, many independent scholarly societies publish their journals through larger publishers. If your institution subscribes to the society journals, it likely does so through the larger publisher. Standardizing the society’s name to the name of the larger publisher is appropriate here, even though it elides some information about the journal. In part for this reason, we recommend below that you keep the publisher information from your original data, simply adding a new column with standardized names.

Name variants can also result from typos. For example, a Web of Science export used in preparing this short course contains about 40 records with the publisher "ACADEMIC PRESS LTD- ELSEVIER SCIENCE LTD" and about 20 with the publisher "ACADEMIC PRESS LTD-ELSEVIER SCIENCE LTD." (Did you spot the difference? It’s the space after the hyphen.) These records were all published by Elsevier’s Academic Press imprint, but a typo means they show up under two different names.

## Sources of inconsistency and inaccuracy in publisher data

Publishers frequently take on titles from other publishers. Entire imprints or publishers can also change hands. A scholarly society may outsource its journals to a larger publisher, move them from one larger publisher to another, or bring them back in house. It’s very difficult to account for this perfectly in data analysis. The different data sources we have described (Web of Science, Scopus, the Lens, and Unpaywall) use different methods for tracking the publisher associated with a title. None of these sources are 100% accurate, but it is important to note that Web of Science, Scopus, and the Lens tend to list the publisher for the title at the time of the item’s publication. Unpaywall uses only the current publisher. For example, for a title that transferred from Wiley to Springer Nature on January 1, 2019, Unpaywall will only report Springer Nature as the publisher, even for documents published before 2019. How much of an issue this is depends, in part, on how far back you are looking. However, using fairly recent data helps, because it reduces the number of changes that will have occurred since the articles were published.

# Procedure

You can do this entire lesson in either OpenRefine or Excel.

You will start with your source dataset, exported from the source you’re working with (e.g., Web of Science, Scopus, or the Lens), which you have built on during the earlier lessons in this course. The work you do in this lesson will be specific to that particular data source, because each source gives publisher information in its own way. If you have not already added Unpaywall data to your source dataset, do so now. You should also make a column for Unpaywall’s "publisher" field by parsing the JSON. (See [Lesson 1.3: Unpaywall API query with OpenRefine](https://docs.google.com/document/d/10sGGXcIUOz6BSL0i7WiKz4amtlVGqsCoeJq5l04ggcY/edit#heading=h.sg2nvk6v3nrh).)

You will also need a list of the major publishers you expect to be represented in your dataset. A good starting point is a list of your library’s major journal subscription contracts. You may be able to get such a list from colleagues working in acquisitions or managing electronic resources. This list can help you think about how to standardize publisher name variants. Of course, do note that your institutions’ researchers are likely publishing with additional publishers your library does not have a relationship with. This is especially true for fully open access publishers, with which your institution would not have a subscription (e.g., PLOS).

If you can’t get such a list, we recommend searching for the publishers mentioned in the appendix to this lesson. To create the list in the appendix, we used large exports from all three data sources containing publishing information from calendar year 2019. We then faceted by publisher in OpenRefine and worked through all the publishers systematically.

This is a laborious process. The more publications in your dataset, the more potential name variants there are, all of which increase the amount of time this lesson will take.

## OpenRefine

### Making your own list of standardized publisher names

1. In OpenRefine, open your source dataset as a project. Using the Custom Tabular Exporter, export and download the column with publisher data. This export option is in the Export menu in the upper right of the OpenRefine interface. De-select all and then select only the column with the publisher data from your original source (not from Unpaywall). For Web of Science data, the column is "PU". For data from Scopus and the Lens, it is "Publisher." Then click on the download tab and select "Tab-separated values" and click Download.
2. Create a new OpenRefine project by importing the TSV you’ve just created. Confirm that the new project has the same number of rows as the full dataset, minus any rows that are blank. We’ll call this new project the "publisher list project."
3. The publisher list project has a list of every publisher name variant that is mentioned in the full dataset. Deduplicate that list, using [OpenRefine: Removing Duplicates, from the University of Illinois Library](https://guides.library.illinois.edu/openrefine/duplicates). As an example, from a file with more than 10,000 records, deduplication reduced the number of publishers to 600.
4. In the publisher list project, click on the arrow menu for the publisher column and choose "Add column based on this column." Name the new column "Standard\_publisher" and leave the expression box as is, containing just "value." This will create a second column with the same values as the first. "Standard\_publisher" is the column you will modify as you clean the data. In this process, you will modify the cells manually, changing the value to the standard publisher name. You will keep all of the original data in the first column.
5. You are now ready to start cleaning this data. We recommend you follow the steps below in the order they are listed. You can keep an eye on your progress by creating a text facet on the "Standard\_publisher" column. This will keep a count of the standard names, which will decrease as you work.

#### Text filter on publisher name

1. Working from the list of publishers you expect to be represented in the dataset (such as the list in the appendix to this lesson), select a publisher.
2. In the publisher list project, set up a text filter on the first column (containing the publisher data from your source dataset), and search for strings from the names of a publisher and its imprints. To identify imprints, look up the publisher in [Sherpa Romeo](https://v2.sherpa.ac.uk/romeo/), or go to the publisher’s own website.
3. Working carefully so as to avoid incorrect standardization, edit the values in the "Standard\_publisher" column to a single standard name. Note that some publishers have misleadingly similar names. In particular, watch out for Springer Publishing Co., which is not part of Springer Nature. To change the values in a batch of records all at once, facet so that only those records are visible. Then, open the triangle menu for the column and select "Edit cells" > "Transform…". Enter the desired new text in the expression box, surrounding it in quotation marks so OpenRefine will treat it as a string. Click OK to change all of the values.
4. Leave the first column unchanged, so you’ll be able to match with the source dataset later.
5. Repeat this process for each of the publishers on your list, including their imprints.

#### Match to Unpaywall publisher field

1. Keeping the publisher list project open, open your source dataset as a project in OpenRefine.
2. In the source project, create text facets on the column containing Unpaywall’s "publisher" field and the column containing publisher data from your original source (Web of Science, Scopus, or the Lens). In the Unpaywall publisher facet, sort by "count" and limit the facet to the first publisher. The other facet, for the publisher field from your original source, may contain more than one entry.
3. Review the additional entries to develop a list of publisher name variants associated with that publisher. Record them in your publisher list project in the "Standard\_publisher" field. For example, if you facet to "Oxford University Press (OUP)" in the Unpaywall publisher data, you may see that your source publisher data contains "Society of American Foresters" in one of those rows. If you research the Society of American Foresters in Sherpa Romeo, you will learn that all of its journals are published by Oxford University Press. You can now go to your publisher list project, search for Society of American Foresters in the first column, and edit the corresponding "Standard\_publisher" cell to say "Oxford University Press." (Why not just fix these in the source project? By recording them in the publisher list project, you’ll be able to apply them later to all the rows in your source project that have that publisher name variant, including rows that do not have DOIs and thus do not contain Unpaywall data.)
4. You have to be a little careful while doing this—in some cases Unpaywall may not match your original data because the journal title has been transferred to a different publisher or Unpaywall is out of date. For example, you might facet to "Wiley" in the Unpaywall publisher data and see that your source publisher data contains "Taylor & Francis Inc" in one of those rows. At time of writing, Taylor and Francis is not part of Wiley, so this is likely a title transfer.

#### Clustering

This technique is especially helpful for cleaning up typos, such as a misspelling in the string at the base of the publisher’s name (e.g., "Elseiver"), or two publisher names that are identical except for a comma. It will not help you identify imprints.

In your publisher list project, in the "Standard\_publisher" column, open the arrow menu and select "Edit cells" > "Cluster and edit…" Initially, there may not be any clusters. You can change the "Method" and "Keying Function" to adjust the algorithms used to cluster values. Work through and merge what you can with various methods and keying functions, being careful not to merge unrelated publishers. For example, some methods will find a match between "CENTERS DISEASE CONTROL & PREVENTION" and "CHINESE CENTER DISEASE CONTROL & PREVENTION," but these are separate publishers that should not be merged. See [OpenRefine: Clustering, from the University of Illinois Library](https://guides.library.illinois.edu/openrefine/clustering) for more information on this technique.

### Mapping standardized publisher names into your dataset

When you have finished work on your publisher list project, return to your source project. Follow [OpenRefine: Joining Projects, from the University of Illinois Library](https://guides.library.illinois.edu/openrefine/joiningprojects) to add a new column named "Standard\_publisher" containing the standardized name corresponding to each publisher name. You will use the first column from your publisher list project, which contains the raw publisher data from your source dataset, as the key for merging the data.

### Analyzing your data

In your source project, you can identify the largest publishers of the institution’s research by creating a text facet on the "Standard\_publisher" column. Choose the option to sort by count to see the publishers with the largest numbers of articles first. If you want to retain or share this data, you can copy it manually into a table or Excel document.

To create a list of a publisher’s journals in which the institution’s authors have published, facet or filter on the publisher field to limit to that publisher. Using the Custom Tabular Exporter, export and download the columns with the journal title and ISSN(s), as TSV. Create a new OpenRefine project by importing the TSV you’ve just created. Deduplicate your list, using [OpenRefine: Removing Duplicates, from the University of Illinois Library](https://guides.library.illinois.edu/openrefine/duplicates). You now have a list of every journal that is mentioned in the full dataset from that publisher.

## Excel

### Making your own list of standardized publisher names

1. Open the source dataset. Create a new sheet. Copy the column with the publisher data to this new sheet. (in Web of Science the column is PU, in Scopus and the Lens it is "Publisher"). Copy the column with the publisher data from your Unpaywall JSONs too. You can work from only one of these lists of publishers but having both can be helpful to identify imprints. For each of the following steps, you will need to record detailed notes as Excel does not keep track of the transformations you make (unlike OpenRefine).
2. Use Excel’s built-in Remove Duplicates function to create a list of the unique names. Highlight the data, then navigate to the "Data" ribbon. Click on the "Remove Duplicates" icon in the "Data Tools" box. Ensure that the "My data has headers" box is checked. You will see that there are checkboxes for which columns to include when deduplicating. Keeping both checked will leave you with a set of values for which there is a unique combination of the publisher from the source data and the Unpaywall publisher. After clicking "OK" you will get a message box reporting how many duplicates were removed and how many values remain. To clean up your view of the data, you can sort it by highlighting both columns and clicking on the "Sort" icon on the "Data" ribbon" (be sure to click on the "My data has headers" box).
3. Now, create a new column to record the standardized publisher name by copying the column with the publisher data from your original dataset. You will make changes to this new column. Rename the column (e.g., "Standard\_publisher") so you remember which column you should be editing.
4. Create a filter over all columns on the sheet. Now start applying filters to the column from the original dataset and modify the name in the "Standard\_publisher" column. Search in the filter in the original list of publisher names for parts of publisher names and enter the standardized name in the newly created column. For example, filtering by searching on "Wiley" can result in names such as Wiley, Wiley-V CH Verlag GMBH, John Wiley & Sons, John Wiley & Sons LTD, Wiley-Blackwell. Change the name to "Wiley" in the Standard\_publisher column. Continue on to combine as many name variants as possible. Referring to the list of major publishers that you expect to be represented in the dataset is useful here.
   1. Note that some publishers have misleadingly similar names. For example, Springer Publishing Co. is not part of Springer Nature.
5. If you still have a large number of publisher names at this point, you can apply filters to the column containing Unpaywall data to identify more imprints. For example, suppose filtering for "Wiley" in the Unpaywall publisher column results in the name "British Veterinary Association" appearing in the original publisher column. Be careful, though, such differences can result from a title transfer rather than "British Veterinary Association" actually being a Wiley imprint. You can use [Sherpa Romeo](https://v2.sherpa.ac.uk/romeo/) to check whether this is truly an imprint. Search on the publisher and see what imprints are listed. Or, if the publisher isn’t in Sherpa Romeo, find the journal title(s) in your original dataset and search in Sherpa Romeo or on the journal’s webpage.
6. Occasionally you will find that the database publisher is completely different from the Unpaywall publisher and it is clearly not an imprint. This is often because of title transfers, as described in the Background section of this lesson. Which version you want to use as the standardized name will depend on what research questions you are trying to answer.

### Mapping standardized publisher names into your dataset

The VLOOKUP function in Excel joins data from one table to another. See [Excel basics](https://docs.google.com/document/d/1ZfGa96mXlEzYL9O-q-QMZYNUINsfdiNRvdsB6HyY_c4/edit#heading=h.qecepd6qt46) for more information on the formula. Note that Vlookup requires that the column with the value you are matching to be to the left of the column with the data you want to join. Depending on how you have set your files, you may need to copy and paste columns to remove empty columns.

You can now use VLOOKUP by adding a new column (you might name it something like "Cleaned Publisher" in the full dataset, instructing Excel to look for the value in the column of the publisher (as downloaded from the database you used) in the new sheet. Be sure to add "FALSE" as the last argument in the VLOOKUP formula, so that you get exact matches.

Remember (referring to [Excel basics](https://docs.google.com/document/d/1ZfGa96mXlEzYL9O-q-QMZYNUINsfdiNRvdsB6HyY_c4/edit#heading=h.qecepd6qt46) if needed) that unless you use IFERROR, you will get an error message ("#N/A") if there is no match. Re-apply filters so that your new column (e.g., "Cleaned Publisher") is included and select only those cells with the error message (or the value you have chosen if you are using IFERROR). Compare these cells with the original data and the Unpaywall data. You will likely have cases where no match was found because there was no value in the original dataset.

You can go one step further to see if there is publisher data from Unpaywall in cases where there was no data from the original source. Keeping the filter on your Cleaned Publisher column, filter out the blanks in the Unpaywall publisher column. The next step is a little tricky. If you copy over the publisher from Unpaywall, but you might be re-introducing names that you had already standardized. Refer to your worksheet containing the publisher data and copy the standardized name into the Cleaned Publisher column.

As always, be careful to keep detailed notes of the changes you have made.

### Analyzing your data

Pivot tables are excellent for this type of analysis. To begin, highlight all of the data in the primary data spreadsheet. Click to add a pivot table in a new sheet and add fields to Filters, Columns, Rows, and Values.

To identify the largest publishers of the institution’s research and how many articles each of them has published during the time period analyzed, you can put the standardized publisher names in the "Rows" box and in "Values" use a field that has an entry in every row in the dataset. The publication ID number in your dataset (Lens ID in the Lens, UT in Web of Science, or EID in Scopus).

To create a list of a single publisher’s journals in which the institution’s authors have published, you can change the field in the "Rows" box to the one that contains the journal title and move the field with standardized publisher data into the "Filters" box.

With a pivot table, you can explore your data further. For example, Some more suggestions:

* Filter on fields like publication type or publication year.
* View more detailed publication patterns by adding fields to "Columns". Publication year (if you have multiple years in your dataset) and OA status can be interesting.

Next lesson: [Lesson 2.5: Estimating APCs using ISSNs](https://docs.google.com/document/d/146HZksAMBm8ev-TyFz9nzmyzAaXvjgRH9EHcxkf7Db8/edit#heading=h.h7t2c2w9oyoe)

# Appendix: Publisher imprints and other names to standardize

This appendix lists just some of the publisher imprints that we regularly come across. You can use these lists to get started consolidating imprints into the larger publisher name.

When searching your data for the imprints and other names listed below, watch out for common abbreviations. For instance, if you search for the string "American Society for Biochemistry and Molecular Biology" but "Society" is abbreviated in your data, you won’t find it. Searching with the terms that are most distinctive and least likely to be abbreviated, or limiting your search to a word stem (as you would if searching with a wildcard character) will help ensure good results.

## Elsevier

Most data sources list at least a few variants with "Elsevier" in their name (e.g., regional subsidiaries) as well as the following entities whose names may not mention "Elsevier", but are part of Elsevier.

* Academic Press
* Acta Materialia
* American Academy of Allergy, Asthma and Immunology
* American College of Allergy, Asthma and Immunology
* American Society for Biochemistry and Molecular Biology
* Association for Behavioral and Cognitive Therapies (formerly Association for Advancement of Behavioral Therapies)
* Association of Veterinary Anaesthetists, the American College of Veterinary Anesthesia and Analgesia and the European College of Veterinary Anaesthesia and Analgesia
* Australian Society for Parasitology
* Bailliere Tindall
* Biophysical Society
* Cell Press
* China University of Mining and Technology
* Chinese Society of Metals
* Chongqing University
* Churchill Livingstone
* CIG Media Group
* Current Biology
* Excerpta Medica
* Institution of Chemical Engineers
* International Association of Great Lakes Research
* JAI Press
* Journal of Materials Science and Technology
* KeAi Publishing
* King Saud bin Abdulaziz University
* Korean Nuclear Society
* Lancet Publishing Group
* Mosby
* Neoplasia Press
* Pergamon Press
* Reed Business Information
* Society for Range Management
* Society for Vascular Surgery
* W.B. Saunders

## Springer Nature

Most data sources list at least a few variants with "Springer" and/or "Nature" in their name (e.g., regional subsidiaries) as well as the following entities whose names may not mention either "Springer" or "Nature." Watch out for [Springer Publishing Company](https://en.wikipedia.org/wiki/Springer_Publishing), which is not part of Springer Nature.

* Adis
* BioMed Central (commonly abbreviated BMC)
* Birkhauser (note there is also a DeGruyter imprint named Birkhauser)
* Central South University of Technology
* Current Medicine Group
* Humana Press
* MAIK Nauka/Interperiodica
* Materials Research Society
* Minerals, Metals and Materials Society
* Palgrave
* Palgrave Macmillan
* Plenum Publishers
* Science in China Press, also called Science China Press
* Steinkopff
* Transaction Publishers, also called Transaction Periodicals Consortium

## Taylor & Francis

Most data sources list at least a few variants with "Taylor & Francis" or "Taylor and Francis" in their name as well as the following entities whose names do not contain those phrases.

* CRC Press
* CRC Press / Balkema
* Dove Medical Press
* Routledge

## Wiley

Wiley has a few subunits with Wiley in their name. It’s also worth checking for the following names, which may not mention Wiley.

* American Society of Agronomy
* Blackwell
* Crop Science Society of America
* Hindawi
* Soil Science Society of America

## De Gruyter

Most mentions of De Gruyter include the string "De Gruyter." It’s also worth checking for the following.

* Birkhäuser (note there is also a Springer imprint named Birkhauser)
* Sciendo

## Wolters Kluwer

You may find Wolters Kluwer publications under the names "Wolters Kluwer," "Lippincott, Williams & Wilkins," "Medknow Publications," or "Ovid Technologies."

## Other publishers

The following publishers may have name variants that need to be cleaned up, but all their name variants include the name and/or abbreviation listed below.

* ACM or Association for Computing Machinery
* AIAA or American Institute of Aeronautics and Astronautics
* American Institute of Physics
* American Psychological Association
* ASTM or American Society for Testing and Materials
* Bloomsbury
* Emerald
* IEEE or Institute of Electrical and Electronics Engineers
* Institute of Physics Publishing (IOP)
* Oxford University Press
* SAGE
* SPE or Society of Petroleum Engineers
* SPIE
* Thieme
* University of Toronto
* World Scientific Publishing