National Institutes of Health
Office of Management

Survey of Bibliometric Resources and Tools

2018 Bibliometrics and Research Assessment Symposium

Candace Norton, NIH Library

Office of Research Services
National Institutes of Health
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

Housekeeping

- This session will be recorded and distributed after editing and captioning.
- Slides will be available following the Symposium.
- Questions are welcome throughout the presentation; please use the microphone.
- Due to the set up of the breakout room, this is a lecture style course and will not be hands on.

Course Objectives

- Identify the most common data sources for use in bibliometric analysis.
- Introduce tools for data cleaning and data visualization.
- Provide resources for future independent investigation.

Data Sources

Web of Science (1)

- Produced by Clarivate Analytics
  - Requires a license to access
- Web of Science Core Collection
  - 69 million article records
  - 1 billion cited references
  - 1900 – present
- Includes 20,000+ scholarly journals in 250+ science, social sciences, and humanities disciplines
- Also includes book data and conference proceedings

Web of Science (2)

- Indexing for each paper includes:
  - all the authors
  - all author affiliations
  - the abstract and keywords (if provided by the author)
  - funding acknowledgements, including agency and grant numbers (if provided)
  - all the cited references
- Additional metadata captured when possible:
  - ORCID identifiers
  - 5,000+ unified institution names

Web of Science (3)

- Analyze Results feature can be used on any set of Web of Science results to group and rank records
- Options include:
  - Publication count by country or institution
  - Who is publishing on a particular topic in a given time frame
- Results are broken down record count and percentage of total results, then sorted or ranked by field
- Analysis can be exported and/or saved

InCites (1)

- Produced by Clarivate Analytics
  - Requires a license to access
- Uses seven editions of Web of Science Core Collection for publication counts and indicators
  - Represent 12,000+ journals, 12,000 annual conferences, 53,000 scholarly books
  - All document types included
  - Date range 1980-2018
  - Covers sciences, social sciences, and humanities

InCites (2)

- InCites Analytics provides access to:
  - People
  - Organizations
  - Regions
  - Journals
  - Conference Proceedings
  - Research Areas
  - Funding Agencies
  - Books
- Can create visualizations and data tables
- Set benchmarks for comparison
- Used to create reports
  - Research Performance
  - Collaborations
  - Custom reports
  - Local Journal Utilization
  - Institution Profiles

Essential Science Indicators

- Produced by Clarivate Analytics
  - Requires a license to access
- Each journal is assigned to one of 22 research fields
  - A journal can only be assigned to one field
  - Multidisciplinary is an option
- ESI reports are calculated on 10 years of data
- ESI Indicators can be viewed in a map view or as a table; citation trends and research fronts also available

Scopus (1)

- Produced by Elsevier
  - Requires a license to access
- Over 71 million records
  - 64+ million post-1970, including references
  - 6.5+ million pre-1970 records, as far back as 1788
- Focus in social sciences, arts and humanities; also includes science, technology, and medicine
- Includes CiteScore metrics, SCImago Journal Rank, Source Normalized Impact per Paper Article-level metrics via PlumX Metrics

Scopus (2)

- Analytical tools
  - Author evaluator feature analyzes an author's publishing output
  - Compare journals feature analyzes journals across multiple metrics
- Search by document, author, affiliation identifier, ORCID
- Updated daily
- API available

SciVal

- Produced by Elsevier
  - Requires a license to access
- 38 million publication records
- 21,915 journals from 5,000 publishers
- Can be used to create visualizations and standardized reports
- Set benchmarks for comparison
- Identify and analyze collaboration opportunities
- Includes topic prominence to evaluate topic momentum

Dimensions (1)

- Produced by Digital Science
  - Basic version available at no cost
  - Enhanced version, Dimensions Plus, available with license
- Updated release in January 2018
- 128 million publications, grants, policy, data and metrics mentions
- Emphasizes connections between research objects, totaling over 4 billion
  - Grants, patents, clinical trials, and altmetrics

Dimensions (2)

- Full text searching available for 50+ million articles and books
  - Includes open access collections and publicly available archives like PubMed Central
- Uses search filters instead of a curated index
- API available
- Visualization, benchmarking, and other analyses available with Dimensions Plus

PubMed

- Produced by the National Library of Medicine
- 28 million citations from Medline, life science journals, and books
- Fields covered include biomedicine, health, portions of life sciences, behavioral sciences, chemical sciences, bioengineering
- No cost to users
- Comprehensive coverage from 1966-present
  - Select coverage of literature prior to 1966
- Does not track citation references

ClinicalTrials.gov

- Maintained by the National Library of Medicine
- Provides access to information on publicly and privately supported clinical studies on human subjects
- Trial records include:
  - Summary information
  - Disease
  - Intervention
  - Funders
  - Protocol
  - Locations
  - Outcomes
  - Adverse Events
- Content available to download for analysis

Other Clinical Trials Sources

- International Clinical Trials Registry Platform (ICTRP)
  - Maintained by the World Health Organization (WHO)
  - Facilitates the registration of the WHO Trial Registration Data Set on all clinical trials, enhancing public accessibility of that information
  - Provides unambiguous trial identification
- U.S. Department of Health & Human Services, Office for Human Research Protections, Listing of Clinical Trial Registries

NIH RePORTER

- Maintained by the NIH Office of Extramural Programs
- Research Portfolio Online Reporting Tools (RePORT) is a website that provides access to reports, data, and analyses of NIH research activities, intra- and extramural
- Useful for tracking publications and patents resulting from work funded by the NIH
- Also captures information from CDC, AHRQ, SAMHSA, and U.S. Department of Veteran Affairs
- ExPORTER available to download data sets for analysis

Federal RePORTER

- Maintained by STAR METRICS, a trademark of U.S. Department of Health and Human Services
- Searchable database of scientific awards from government agencies
- Users can search across
  - Agencies
  - Fiscal years
  - Award's project leader
  - Keyword search of project title, terms, or abstracts
- Federal ExPORTER available to download data sets for analysis

iCite

- Produced by the NIH Office of Portfolio Analysis
  - Freely available to users
- Web application that provides a panel of bibliometric information for journal publications within a defined analysis group
  - Total number of articles within the analysis group
  - Mean number of articles published per year
  - Number of citations for articles in the analysis group per year
  - Relative Citation Ratio (RCR)
  - Weighted RCR: the sum of the RCRs for the articles within the analysis group
- Limited to analyzing only articles that appear in PubMed

Patents

- U.S. Patent and Trademark Office
  - USPTO Patent Full-Text and Image Database (PatFT)
  - USPTO Patent Application Full-Text and Image Database (AppFT)
- European Patent Office
  - Espacenet Patent search
- Canadian Patent Database
- German Patent and Trademark Office (GPTO)
  - DEPATISnet
- World Intellectual Property Organization
  - PATENTSCOPE Search Service
- Japan Patent Office
- United Kingdom Intellectual Property Office (IPO)

Data Cleaning

Excel (1)

- Excel is a program for creating, processing, analyzing, and visualizing tables of data
- Easy to enter and view data
- Formulas available to calculate data
  - Includes basic arithmetic to more complicated functions
- May not be the easiest tool for resolving messier data issues
- Importing data from alternate file formats (.csv instead of .xls or .xlsx) can be tricky

Excel (2)

- Ways to use Excel to clean messy data
  - Inserting and deleting rows and columns
  - Creating tables
  - Merging and splitting data into multiple columns
  - Splitting and combining data in columns using Flash Fill
  - Removing duplicate records
  - Correcting dates and times
  - Changing the case of the text

OpenRefine (1)

- OpenRefine is a powerful tool for working with messy data
- OpenRefine can clean data, transform data from one format to another, and can extend data via web services and external data
- OpenRefine uses a selection of algorithms to cluster and merge values to facilitate name disambiguation
- OpenRefine is available for download online at http://openrefine.org/download.html

OpenRefine (2)

- Uses rows and records when working with data
  - OpenRefine can link together multiple rows as belonging to the same Record
- Resolve inconsistencies in a data set
- Helps split data up into more granular parts
  - splitting up cells with multiple authors into separate cells
- Match local data up to other data sets
- Enhance a data set with data from other sources

VOSviewer

- Produced by CWTS, Leiden University
  - Available at no cost to users
- Software tool for constructing and visualizing bibliometric networks
- Accepts data imports from WoS, Scopus, Pubmed, CrossRef, and RIS format
- Thesaurus files can be constructed to facilitate data cleaning using two columns, *label* and *replace by*
  - Thesaurus files must be created by the user
  - Can be used to ignore or disregard irrelevant terms or conduct name cleaning

R and RStudio

- R is an open source programming language for statistical computing and graphics
- R Studio is an integrated development environment (IDE)
- R has many packages (collection of functions) available to work with messy data
  - tidyr
  - dplyr
  - lubridate
  - readr
  - stringr
  - magrittr

GRID

- Produced by Digital Science
  - Freely available for download
- Provides an automatic disambiguation service to algorithmically match author affiliation strings to institutions
  - Location aware disambiguation
  - Multilingual names
  - Name variants
- Available at https://grid.ac/

Science of Science (Sci2)

- Designed by Indiana University
  - Freely available to users
- Used for creating temporal, geospatial, topical, and network analysis, and visualizations
- Able to clean, analyze, and visualize a variety of data formats with built-in algorithms available in a point and click interface
- Supports data import in a variety of formats
- Allows users to save workflows for reproducibility

Data Visualization

Excel

- Excel is a program for creating, processing, analyzing, and visualizing tables of data
- Comes with a number of preset chart options available
  - Easily preview different chart types
  - Quickly adjust design elements
- Access to add-ins can increase Excel's capabilities
  - Bubbles
  - GIGRAPH
  - Power Map
  - People Graph
- Many resources available online to learn advanced tips for data visualization in Excel

Gephi

- Open-source software for network visualization and analysis
- No programming skills required to use
- Useful for generating collaboration networks, topic analysis graphs
- Provides many options for layout algorithms
- Also includes community detection algorithms, and the ability to embed a timeline to visualize a graph over time
- Easily exports to a variety of files types

VOSviewer

- Produced by CWTS
  - Available at no cost to users
- Software tool for constructing and visualizing bibliometric networks
  - Networks can include journal, researchers, publications
  - Citation, bibliographic coupling, co-citation, co-authorship relations
  - Text mining functionality for word co-occurrence networks
- Accepts data imports from WoS, Scopus, Pubmed, CrossRef, and RIS format

Cytoscape

- Open source software platform for complex network analysis and visualization
- Originally designed for use with molecular and genetic data, but has since expanded
- Variety of layout algorithms available
- VizMapper feature is fairly intuitive to use to customize the details of the network visualization
- Free training resources available online

Tableau

- Interactive data visualization and analytics tool known for its use in creating dashboards
- Very flexible in the data sources it can accept and visualize
- Tiered access model available with several low or no cost options
  - Tableau Public
  - Tableau Academic
  - Tableau Desktop
  - Tableau Online
  - Tableau Server
- Many free training resources available

R and RStudio

- R is an open source programming language for statistical computing and graphics
- R Studio is an integrated development environment (IDE)
- R has several packages (collections of R functions) that create high quality graphics
  - ggplot2
  - igraph
  - RColorBrewer
  - plotly
  - lattice
  - ggmap

Inkscape

- Open-source vector graphics editor that uses Scalable Vector Graphics (SVG) as the native format
  - Similar to Adobe Illustrator, Corel Draw, Freehand, or Xara X
- Most often used to create graph legends and tidy label placement on network graphs
- Can be used for original design work
- Many tutorials and learning resources available

Additional Resources

Books (1)

- Sugimoto C., Larivière V. Measuring Research: What Everyone Needs to Know. Oxford University Press; 2017.
- Roemer R., Borchardt, R. Meaningful Metrics: A 21st Century Librarian's Guide to Bibliometrics, Altmetrics, and Research Impact. The Association of College & Research Libraries; 2015.
- Börner K. Visual Insights: A Practical Guide to Making Sense of Data. MIT Press; 2014.
- Börner K. Atlas of Science: Visualizing What We Know. MIT Press; 2010.

Books (2)

- Börner K. Atlas of knowledge: anyone can map. MIT Press; 2015.
- Ding, Y., Rousseau, R., & Wolfram, D. (Eds.). Measuring Scholarly Impact: Methods and Practice. Springer; 2014.
- Barabási, A. Network Science. Cambridge University Press; 2015.
- Rousseau, R., Egghe, L., Guns, R. Becoming Metric-Wise: A Bibliometric Guide for Researchers. Chandos Publishing; 2018.
- Cronin, B., Suigmoto, C. (Eds.) Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. MIT Press; 2014.

Articles (1)

- Borner K, Sanyal S, Vespignani A. Network science. *Annu Rev Inf Sci Technol*. 2007;41:537-607.
- Newman MEJ. The structure of scientific collaboration networks. *Proc Natl Acad Sci U S A*. 2001;98(2):404-9.
- Barabasi AL. Scale-Free Networks: A Decade and Beyond. *Science*. 2009;325(5939):412-3.
- Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391.

Articles (2)

- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.
- de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation, 25*(2), 161-169.
- Corrall, S., Kennan, M. A., & Waseem, A. (2013). Bibliometrics and research data management services: emerging trends in library support for research. *Library Trends, 61*(3), 636-674.

Journals

- Journal of Informetrics (https://www.sciencedirect.com/journal/journal-of-Informetrics)
- Journal of Scientometric Research (http://www.jscires.org/)
- Journal of the American Society for Information Science and Technology (https://onlinelibrary.wiley.com/journal/15322890)
- Journal of the Medical Library Association (http://jmla.mlanet.org/ojs/jmla)

Blogs

- The Bibliomagician
- CWTS Blog
- Leiden Manifesto Blog

Twitter

- @LudoWaltman
- @LizzieGadd
- @cwtsleiden
- @lutzbornmann
- @neesjanvaneck
- @Ednoyons
- @paulwouters
- @sig_met
- @stefhaustein
- @lariviev
- @scimago
- @Scopus
- @ResMetrics
- @DORAssessment
- @altmetric
- @csugimoto
- @timothydbowman
- @jevinwest

Listservs and Groups

- LIS-Bibltiometrics list
- SIGMetrics list
- ScholComm list
- Research Impact Services Google Group

Online Resources

- Metrics-Toolkit.org
  - Online tool providing evidence-based information about various research metrics across disciplines
- ResponsibleMetrics.org
  - Forum for discussion built on the UK's Independent Review of the Role of Metrics in Research Assessment and Management
- LeidenManifesto.org
  - Guidelines outlining 10 principles to guide research evaluation

Conferences and Trainings

- NIH Library and SLA MD Bibliometrics and Research Assessment Symposium
- Week-long training at Centre for Science and Technology Studies (CWTS) at Leiden University
  - Partnering with Galter Health Sciences Library at Northwestern University to offer training in the US
- CE sessions at major library conferences
  - American Library Association (ALA)
  - Special Libraries Association (SLA)
  - Medical Libraries Association (MLA)
- Metrics SIG meeting at ASIST

Questions?

Email: Candace.Norton@nih.gov

Visit: www.nihlibrary.nih.gov/bibliometrics


DEPARTMENT OF HEALTH & HUMAN SERVICES · USA


NIH
National Institutes of Health
Office of Management