Automated Data Collection & Extraction

Tools & Resources

Optical Character Recognition

1. Video Primer: What is OCR?
2. Online OCR Tools - text and data from pdfs into csv,txt, etc.
   - Google Drive - Video Primer: OCR with Google Drive
   - Free online sites for small batches (can upgrade for larger numbers of files)
     - Free Online OCR 1
     - New OCR
     - pdf to excel
     - OnlineOCR
     - PDFTables will convert PDF to .csv, and has an API so you can do your conversions in bulk with R. You can do ~25 pages free; large numbers are reasonably priced.
   - Mathpix Snip digitizes handwritten or printed text, and copies outputs to the clipboard that can be pasted into LaTeX editors like Overleaf, Markdown editors like Typora, Microsoft Word, and more.
3. R tools for OCR: `pdfreader`, `tabulizer`, and `pdfreader`
   - R package `pdftools`
   - written tutorial 1
   - written tutorial 2
   - written tutorial 3
   - Video Tutorial 1
   - Video Tutorial 2
   - Detailed Blog Post / Tutorial
   - Convert PDF to text in R OCR pdftools
   - PDFtools in R
4. Extracting Tables from images using R package `magick`.
   - Detailed Blog Post / Tutorial

Extracting Data from Published Figures

1. Ankit Rohagni's Web Plot Digitizer
   - WPD Video Tutorial
   - WPD Tutorial Blog Post
2. Alternative 1: R package `magick`
3. Alternative 2: GetData extracts data automatically from scanned images (~$30).

4. Alternative 3: R package `digitize` will extract data from scatterplots within the R environment. This article will walk you through the process.

Text Mining

1. Text Mining with R by Julia Silge and David Robinson
2. `gutenbergr`: Download and Process Public Domain Works from Project Gutenberg. Tutorial can be found here
3. Atanassova I, Bertin M and Mayr P (2019) Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. Front. Res. Metr. Anal. 4:2. doi: 10.3389/frma.2019.00002
4. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S (2018) A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol 14(2): e1005962. https://doi.org/10.1371/journal.pcbi.1005962
5. Salloum, S.A., Al-Emran, M., Monem, A.A., Shaalan, K. (2018). Using Text Mining Techniques for Extracting Information from Research Articles. In: Shaalan, K., Hassanien, A., Tolba, F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, vol 740. Springer, Cham. https://doi.org/10.1007/978-3-319-67056-0_18
6. Simon, C., Davidsen, K., Hansen, C. et al. BioReader: a text mining tool for performing classification of biomedical literature. BMC Bioinformatics 19, 57 (2019). https://doi.org/10.1186/s12859-019-2607-x Extracting Body Text from Academic PDF Documents for Text Mining
7. Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text mining methodologies with R: An application to central bank texts. Machine Learning with Applications, 8, 100286.link
8. Yu, C., Zhang, C., & Wang, J. (2020). Extracting Body Text from Academic PDF Documents for Text Mining. arXiv preprint arXiv:2010.12647.
9. Gulo, C. A., & Rúbio, T. R. (2015, January). Text Mining Scientific Articles using the R. In Doctoral Symposium in Informatics Engineering. linl
10. jstor: An R package for Analysing Scientific Articles
11. Processing Records from the WOS and Scopus for Text Mining & Biliometrics
    - R package `refsplitr`
    - R package `bibliometrix`

Scraping information from websites

1. Noortje Marres & Esther Weltevrede (2013) Scraping the Social?, Journal of Cultural Economy, 6:3, 313-335, DOI: 10.1080/17530350.2013.772070

2. Library Carpentry Lesson on Webscraping
3. Start Here: Introduction to webscraping
4. Video: Scraping WebData in R with rvest
5. Video: Practical Introduction to Web Scraping using R
6. Very nice written tutorial…
7. ….and another one, this time from the UC Business Analytics R Programming Guide
8. scraping HTML text and scraping HTML tables
9. SelectorGadget is useful to id CSS selectors

Social Media Data

1. Scraping Twitter Data with R or with Tweetsets

Cell Phone Data

1. Exploratory analyses Part 1 and Part 2

Automated Image Analysis

1. Pennekamp, F. and Schtickzelle, N. (2013), Implementing image analysis in laboratory-based experimental systems for ecology and evolution: a hands-on guide. Methods Ecol Evol, 4: 483-492. https://doi.org/10.1111/2041-210X.12036
2. How to build your own image recognition app with R! Part 1 and Part 2

Wearable Devices & RFID tags

1. What is an RFID tag?
2. Rafiq, K., Appleby, R. G., Edgar, J. P., Radford, C., Smith, B. P., Jordan, N. R., Dexter, C. E., Jones, D. N., Blacker, A. R. F., & Cochrane, M. (2021). WildWID: An open-source active RFID system for wildlife research. Methods in Ecology and Evolution, 12, 1580– 1587. https://doi.org/10.1111/2041-210X.13651
3. Build your own RFID device
4. Izmailova, E.S., Wagner, J.A. and Perakslis, E.D. (2018), Wearable Devices in Clinical Trials: Hype and Hypothesis. Clin. Pharmacol. Ther., 104: 42-52. https://doi.org/10.1002/cpt.966
5. Loncar-Turukalo T, Zdravevski E, Machado da Silva J, Chouvarda I, Trajkovik V. Literature on Wearable Technology for Connected Health: Scoping Review of Research Trends, Advances, and Barriers J Med Internet Res 2019;21(9):e14017 doi: 10.2196/14017
6. Why Should Sociologists Care about Wearable Tech?

7. Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. Perspectives on Psychological Science 11(6), 838-854 link
8. Seifert Alexander, Hofer Matthias, Allemand Mathias. 2018. Mobile Data Collection: Smart, but Not (Yet) Smart Enough. 12. Frontiers in Neuroscience https://www.frontiersin.org/article/10.3389/fnins.2018.00971

Build your own automated data collection tool, Rasberry PI

1. Calipers that dump data directly to Excel link
2. PiSpy: An Affordable, Accessible, and Flexible Imaging Platform for the Automated Observation of Organismal Biology and Behavior
3. Jolles, J. W. (2021). Broad-scale applications of the Raspberry Pi: A review and guide for biologists. Methods in Ecology and Evolution, 12, 1562– 1579. https://doi.org/10.1111/2041-210X.13652

Databases and Websites from which you can extract data

1. Government
   - Data.gov (the open data portal of the US Government) and Using Data.gov APIs in R
   - the rOpengov Project
   - Open Fiscal Data Package
   - `educationdata`: Retrieve data from the Urban Institute's Education Data API as a data.frame for easy analysis. See also here
   - a huge list of data sources for social scientists available with R tools
   - accessing World bank Data with R
2. US & World Census Data
   - A Guide to Working with US Census Data in R
   - R Package `tidycensus`
   - Tutorial 1
   - Tutorial 2
   - R package `ipumsr`: The ipumsr package helps import IPUMS extracts from the IPUMS website into R. IPUMS provides census and survey data from around the world integrated across time and space.
3. Education Data
   - `edbuildr`: import EdBuild's master dataset of school district finance, student demographics, and community economic indicators for every school district in the United States.

   - Building R and Stata packages for the Education Data Portal

4. Other Online Data Sets & Tools for Accessing them
    - Giant compendium of open datasets #1
    - Data on Amazonia

    - R package `bdc`: toolkit for gathering & cleaning biodiversity data
    - EcoRetriever: automates the tasks of finding, downloading, and cleaning up publicly available ecological data, and then stores them in a local database or csv files.
    - litsearcher an R package to facilitate quasi-automatic search strategy development for systematic review
5. Correia, R.A., Ladle, R., Jarić, I., Malhado, A.C.M., Mittermeier, J.C., Roll, U., Soriano-Redondo, A., Veríssimo, D., Fink, C., Hausmann, A., Guedes-Santos, J., Vardi, R. and Di Minin, E. (2021), Digital data sources and methods for conservation culturomics. Conservation Biology, 35: 398-411. https://doi.org/10.1111/cobi.13706