

# Automated Data Collection & Extraction

## Tools & Resources

### Optical Character Recognition

1. Video Primer: [What is OCR?](#)
2. Online OCR Tools - text and data from .pdf 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.
  - [Amazon TextExtract](#)
  - [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`
  - R package `pdftools`
  - R package `tabulapdf`
  - [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](#)
  - More advanced but more powerful from the Programming Historian: [OCR with Google Vision API and Tesseract](#)
4. Extracting Tables from images using R package `magick`.

- R package [magick](#) (*this package actually includes several very powerful tools for image processing; this is just one of the things you can do with it*)
- 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](#)

## Other Useful Text Mining Literature

- \* Atanassova I, Bertin M and Mayr P (2019) Editorial: Mining Scientific Papers: NLP-enhanced
  - \* Westergaard D, Størfeldt H-H, Tønnsberg C, Jensen LJ, Brunak S (2018) A comprehensive approach to text mining of scientific publications
  - \* Salloum, S.A., Al-Emran, M., Monem, A.A., Shaalan, K. (2018). Using Text Mining Techniques to Analyze Scientific Papers
  - \* Simon, C., Davidsen, K., Hansen, C. et al. BioReader: a text mining tool for performing text mining on biological text
- ### Extracting Body Text from Academic PDF Documents for Text Mining
- \* Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text mining methodologies with R: An overview
  - \* Yu, C., Zhang, C., & Wang, J. (2020). Extracting Body Text from Academic PDF Documents for Text Mining
  - \* Gulo, C. A., & Rúbio, T. R. (2015, January). Text Mining Scientific Articles using the

1. Processing and Analyzing Records from the Web of Science and Scopus for Text Mining & Bibliometrics
  - R package [refsplitr](#)
  - R package [bibliometrix](#)
  - [jstor](#): An R package for Analysing Scientific Articles

## Scraping information from websites

1. Library Carpentry [Lesson on Webscraping](#)
2. Start Here: [Introduction to webscraping](#)
3. Video: [Scraping WebData in R with rvest](#)
4. Video: [Practical Introduction to Web Scraping using R](#)
5. Very nice [written tutorial](#)...
6. ....and another one, this time from the [UC Business Analytics R Programming Guide](#)
7. [scraping HTML text](#) and [scraping HTML tables](#)
8. [SelectorGadget](#) is useful to id CSS selectors.
9. Noortje Marres & Esther Weltevrede (2013) Scraping the Social?, Journal of Cultural Economy, 6:3, 313-335, DOI: 10.1080/17530350.2013.772070

## Cell Phone Data

1. Exploratory analyses [Part 1](#) and [Part 2](#)

## Social Media (Facebook, Flickr, etc,)

1. How to extract [Biodiversity Data from Facebook](#)
2. Fox, Nathan, Tom August, Francesca Mancini, Katherine E. Parks, Felix Eigenbrod, James M. Bullock, Louis Sutter, and Laura J. Graham. ““photosearcher” package in R: An accessible and reproducible method for harvesting large datasets from Flickr.” SoftwareX 12 (2020): 100624. <https://www.sciencedirect.com/science/article/pii/S235271102030337X>

## 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](https://doi.org/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, Raspberry 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

Overview: 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>

### Government

- \* [Data.gov] (<https://www.data.gov/>) (the open data portal of the US Government) and [Usi
- \* the [rOpengov Project] (<http://ropengov.org/>)
- \* [Open Fiscal Data Package] (<http://www.fiscaltransparency.net/ofdp/>)
- \* ['educationdata'] (<https://urbaninstitute.github.io/education-data-package-r/#education>)
- \* a [huge list] (<https://cengel.github.io/gearup2016/SULdataAccess.html>) of data sources
- \* accessing [World bank Data] (<https://vincentarelbundock.github.io/WDI/>) with R

### US & World Census Data

- \* [A Guide to Working with US Census Data in R] (<https://rconsortium.github.io/censusguid>)
- \* R Package ['tidycensus'] (<https://cran.r-project.org/web/packages/tidycensus/tidycensus>)
- \* [Tutorial 1] (<https://walkerke.github.io/tidycensus/articles/basic-usage.html>)
- \* [Tutorial 2] (<http://www.computerworld.com/article/3120415/data-analytics/how-to-downlo>)

\* R package `[‘ipumsr’]` (<https://tech.popdata.org/ipumsr/>): The ipumsr package helps impor

### Education Data

\* `[‘edbuildr’]` (<https://github.com/EdBuild/edbuildr>): import [EdBuild’s master dataset] (h

\* [Building R and Stata packages for the Education Data Portal] (<https://urban-institute.>

### Other Online Data Sets & Tools for Accessing them

\* Giant [compendium of open datasets #1] (<https://github.com/awesomedata/awesome-public-d>

\* [Data on Amazonia] (<https://datazoomamazonia.com.br/>)

\* 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