

FINDING DATA ON THE WEB

Local voices. Global change.

A very common data format ...



Internews
Local voices. Global change.

Where is Data?

Data is Collected Every Time Something is:

- Registered
- Inspected
- Purchased
- Voted on
- Surveyed

Finding Data on the Web

- Understanding Data Formats
- Advanced Google Searches
- Scraping data from the web
- Exercise
- Summary & Resources

Finding Data on the Web:

Data Formats for Download

- Portable Document Format (PDF): charts that contain data but are saved in a unified document with text
- Excel file (xls): data is saved as a table readable by Microsoft Excel
- Comma separated values (CSV): Plain text file with each data separated by a comma



Finding Data on the Web

- **Understanding Data Formats**
- Exercise
- Advanced Google Searches
- Scraping data from the web
- Exercise
- Summary & Resources

Contents

Understanding Data Formats

- Machine readable and unstructured
- Scanned and computer-generated
- Exercise
- Converting scanned images into useable formats
- Exercise
- Converting unstructured formats into machine readable formats



Contents

Understanding Data Formats

- **Machine readable and unstructured**
- Scanned and computer-generated
- Exercise
- Converting scanned images into useable formats
- Exercise
- Converting unstructured formats into machine readable formats

Summary & Resources

Understanding Data Formats

Machine readable

CSV (comma-separated values) or
TSV (tab-separated values), Excel
(.xls)

Unstructured

(PDF, Word) and bitmap images
(GIF, JPEG, PNG, BMP)



Internews
Local voices. Global change.

Understanding Data Formats: Machine Readable

pollingcenters_2009 - Excel

Province	District ID	Polling Ce	PC Code	Estimated	Total PS	Male	Female	Kuchi
Kabul	101	Qool Bahc	101510	3000	6	0	0	6
Kabul	101	taalab jaan	101449	1000	6	0	0	6
Kabul	101	dasht kha	101448	6000	15	0	0	15
Kabul	101	haaji satar	101331	2900	5	0	0	5
Kabul	101	malwari g	101509	3000	6	0	0	6
Kabul	101	religiose s	101473	4200	8	0	0	8
Kabul	101	Spin ghar	101511	2500	6	0	0	6
Kabul	101	haaji gulw	101332	2900	5	0	0	5
Kabul	101	bala joy a	101014	1000	2	1	1	0
Kabul	101	maiwand	101003	1500	3	2	1	0
Kabul	101	sar gardan	101012	600	2	1	1	0
Kabul	101	cotton sel	101005	2000	4	2	2	0
Kabul	101	Takya kha	101011	1500	3	2	1	0
Kabul	101	Imam Bac	101017	1500	3	2	1	0
Kabul	101	Achakzaye	101001	3800	7	4	3	0
Kabul	101	sedokan c	101004	3800	7	5	2	0
Kabul	101	shor bazan	101002	2400	4	2	2	0
Kabul	101	babay kha	101007	4300	8	5	3	0
Kabul	101	eid gah,ei	101008	4800	8	6	2	0
Kabul	101	bagh qazi,	101009	2400	4	2	2	0
Kabul	101	baghban k	101015	1500	3	2	1	0
Kabul	101	Ansuri Ba	101016	1500	3	2	1	0
Kabul	101	Jafaria Ma	101013	3800	7	5	2	0
Kabul	101	pol bagh a	101010	3800	7	5	2	0
Kabul	101	bala qalag	101018	1500	3	2	1	0
Kabul	101	Baihaqi H	101006	1500	3	2	1	0
Kabul	101	qala fateh	101208	2700	5	3	2	0
Kabul	101	qala cham	101209	1000	2	1	1	0
Kabul	101	Chamand	101201	1500	3	2	1	0



Internews
Local voices. Global change.

Understanding Data Formats

Unstructured

Land Quest Media Matters - Word

Four continents, four languages, four journalists

Internews in Kenya has learned that producing data journalism in Kenya requires a data community, even if that community spans four continents.

Land Quest, an experiment in cross-border investigative journalism by two European, two Kenyan and one American journalist seeks to redefine both the focus and the audience of development reporting. The technical platform was built by Internews partner ((o))Ecolab, an initiative to create and transform journalism practices for reporting on the environment based in Brazil.

The data reveals Kenya as the battlefield between two competing financial interests: the flow of aid money from Europe to Kenya and multinational profits from Kenya to the Europe Union, Kenya's second largest trading partner after China. We wanted a simple way for Kenyans and global citizens to be able to see aid money flows into Kenya to help strengthen institutions and private companies, from agro-industrialists to oil barons, profit from unregulated resources flowing back to Europe.



Internews
Local voices. Global change.

Contents

Understanding Data Formats

- Machine readable and unstructured
- **Scanned and computer-generated**
- Exercise
- Converting scanned images into useable formats
- Exercise
- Converting unstructured formats into machine readable formats

Understanding Data Formats

PDFs

Scanned Images
(Unstructured, not
searchable)

This is a scanned image of a Spanish tax form. It contains various fields for personal data, tax information, and a table for declaring assets and income. The text is in Spanish and the form is structured with boxes and lines for data entry.

Computer-generated
(searchable)

Tables (Structured,
searchable)

This is a computer-generated table with multiple rows and columns. The data is structured and searchable. The table appears to be a financial or statistical report, with columns for categories and rows for individual data points.

Complex formats
(Unstructured,
searchable)

This is a complex computer-generated form with multiple sections and fields. It contains various text boxes, checkboxes, and structured data sections. The form is designed for data entry and is searchable.

Understanding Data Formats: Machine-generated PDFs

Copying and pasting into Excel does not usually work

Free online tools for data in English:

- cometdocs.com
- pdftoexcelonline.com
- zamzar.com
- pdftoexcel.org

Free desktop software

tabula.nerdpower.org



Internews
Local voices. Global change.

Contents

Understanding Data Formats

- Machine readable and unstructured
- Scanned and computer-generated
- **Exercise**
- Converting scanned images into useable formats
- Exercise
- Converting unstructured formats into machine readable formats

Understanding Data Formats: Exercise

1. Open the File: [UNDP Human Development Report Sri Lanka 2013](#)
2. Go to [www.zamzar.com](#)
3. Upload your file
4. Select “Convert to” xlx.
5. Enter your e-mail address
6. You should receive an e-mail when your document is ready
7. Also try:
cometdocs.com
pdfstoexcelonline.com
zamzar.com
pdfstoexcel.org



Contents

Understanding Data Formats

- Machine readable and unstructured
- Scanned and computer-generated
- Exercise
- **Converting scanned images into useable formats**
- Exercise
- Converting unstructured formats into machine readable formats

Understanding Data Formats: Scanned Images

Optical Character Recognition Software

- For documents that you can't search or select text
- Quality depends on quality of image, clarity of text
- Adobe Acrobat Professional: paid
- Google Docs: free
- [Document Cloud](#): free. Advantage is that it allows you to annotate and publish your document as part of a multimedia publication. Your media outlet must request an account.

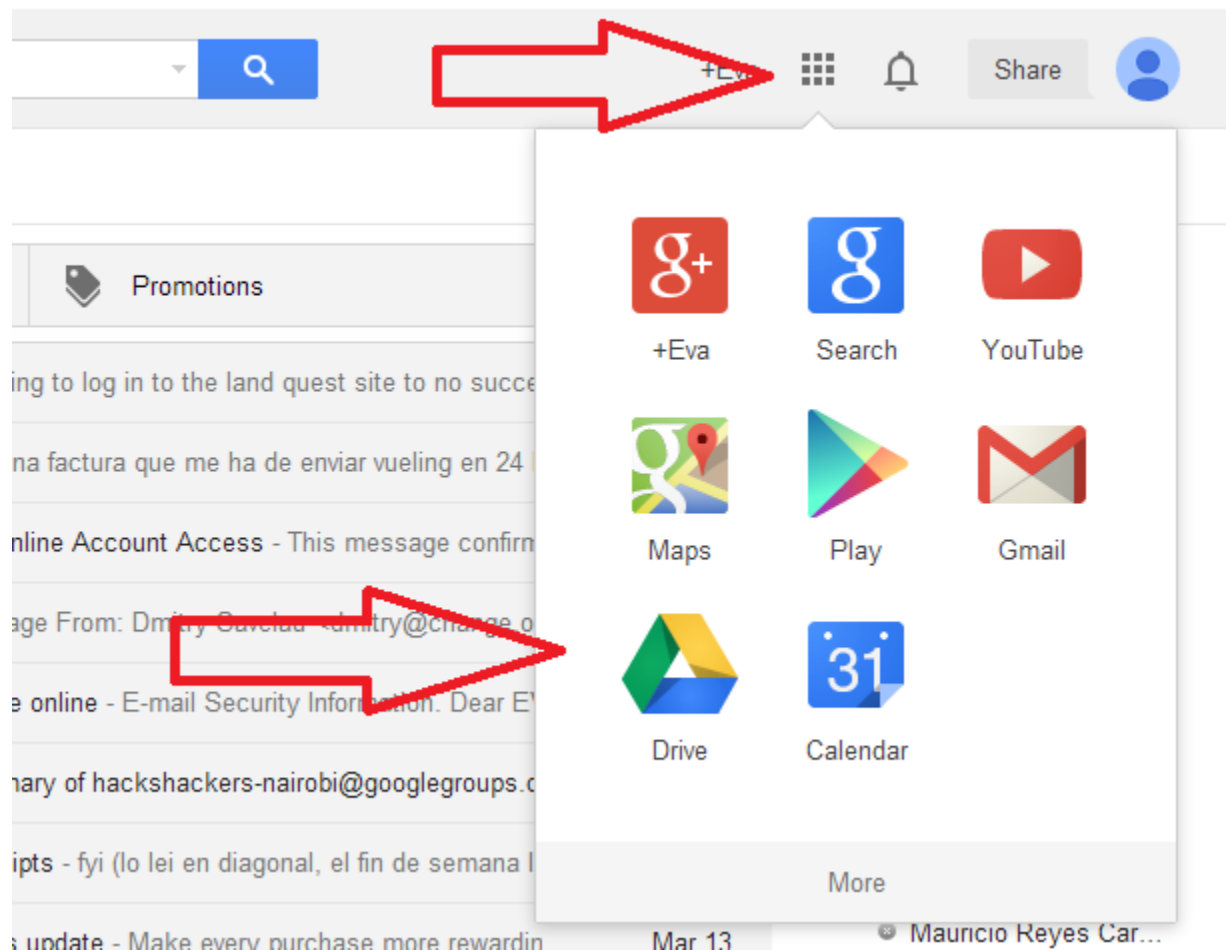
Contents

Understanding Data Formats

- Machine readable and unstructured
- Scanned and computer-generated
- Exercise
- Converting scanned images into useable formats
- **Exercise**
- Converting unstructured formats into machine readable formats

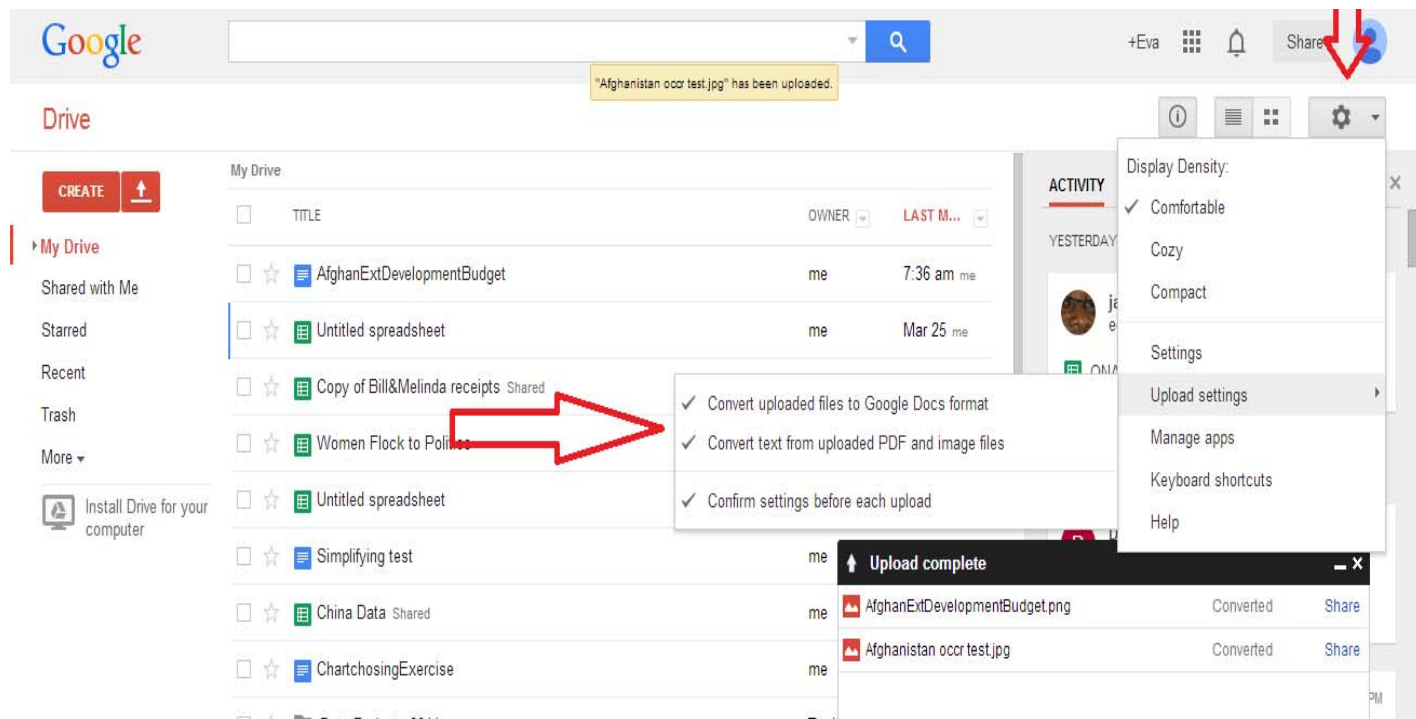
Understanding Data Formats: Scanned Images Exercise

1. Open and save [Bank Reconciliation Statement](#)
2. Open Google Drive (click on grid in the top right corner of G-mail and select Drive)



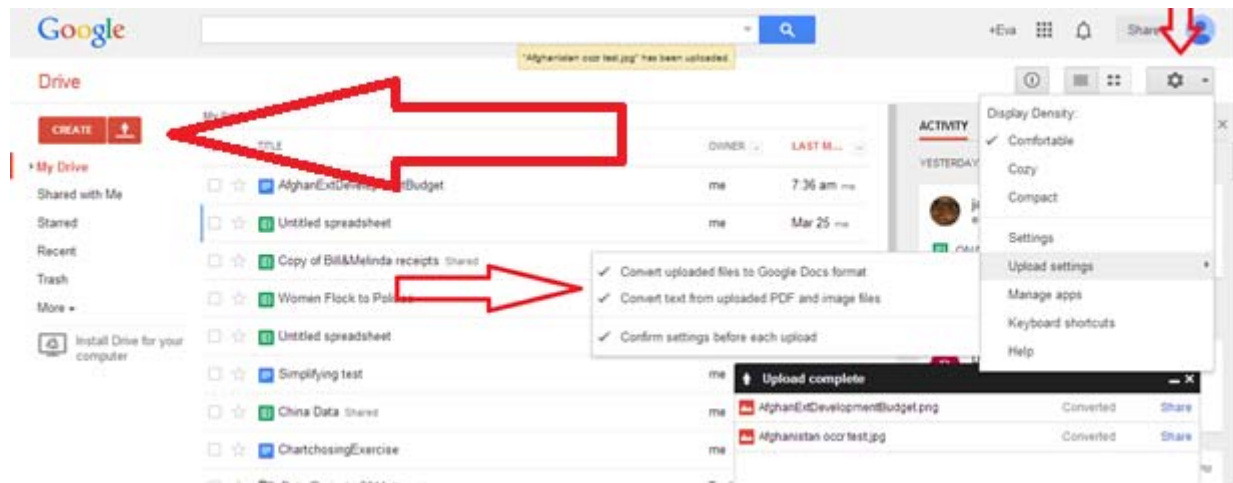
Understanding Data Formats: Scanned Images Exercise

3. Go to Setting and Upload Settings: Select all options



Understanding Data Formats: Scanned Images Exercise

4. Next to the Create button, click on the Upload button
5. Select your PDF or Image file
6. Once the file is uploaded it should appear on your drive with both the image and any text that Google was able to extract.



Contents

Understanding Data Formats

- Machine readable and unstructured
- Scanned and computer-generated
- Exercise
- Converting scanned images into useable formats
- Exercise
- **Converting unstructured formats into machine readable formats**

Understanding Data Formats: Unstructured

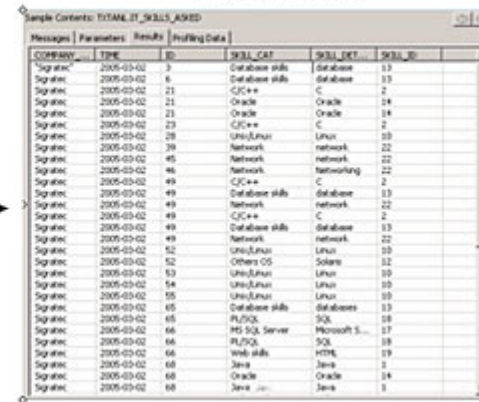
- Columns, data divided among various lines, mixed with text
- Automatic tools probably won't work
- The PDFs will have to be extracted and a programmer will use programming language to write a script to extract the data that you want

Understanding Data Formats: Unstructured

Unstructured text

work smarter. Transforming our business models. Driving our competitive cost structure. Setting up end-to-end ownership. And making sure we make the most of our resources. IT creates the re-usable building blocks with which we assemble integrated solutions, and deliver the consulting that makes the technology work. Excellence is the minimum standard acceptable job description. Developer position with focus on MS Access application with components in MS Access, SQL Server, ASP, and ASP .Net, Java. Activities will include but not limited to the followings: Develop/enhance systems base on requirements and analysis provided by customers; Perform Design, Coding, Unit Testing, and Documentation on assigned work; Provide guidance to other developer/testers in the same organization as needed; Resolve technical issues for customers with quick response and high quality; Attending technical and project management meetings; Help on enhancement/maintenance on legacy components; Communicate properly to both on-shore and off-shore partners at different time zones. Qualifications: The ideal candidate for this position will be someone who is a developer on Tera platforms and products. The candidate should have been through large and small projects and have worked with the full IT Software Development Life Cycle. The candidate should understand that design, unit test, documentation part of the development effort. The candidate should also be flexible to help on enhancement/maintenance on legacy components that are written in Pro*C, PL/SQL, Oracle. Required Experience: Must have 2 or 3+ years of experiences on MS Access development along with other required skill sets listed below. Must have 2+ years of experience in ASP and ASP .NET. Should be able to work independently with minimum supervision and have the ability to use his/her analytical skills to solve problems. Required Skills & Training: Fluency in MS Access VBA development. Fluency in ASP, ASP .Net. Versatile in SQL Server 2000 development tool set including DTS, BCP, stored procedures, TSQL, etc. Versatile in DOS shell scripting. Required Availability & Location: This is an off shore position in Shanghai, China. The candidate will work with the development team in Houston, USA. The candidate must be available for occasional evening Central US time phone conference with Houston, USA to resolve issues or gather requirements. Additional Desired Experience, Skills, and Training: Informatica, Oracle, Business Objects, C, C++, Java, JSP.

Structured text



COMPANY	TIME	ID	SKILL_CAT	SKILL_DET	SKILL_ID
Signatix	2005-03-02	3	Database skills	Database	13
Signatix	2005-03-02	6	Database skills	Database	13
Signatix	2005-03-02	21	C/C++	C	2
Signatix	2005-03-02	21	Oracle	Oracle	14
Signatix	2005-03-02	21	Oracle	Oracle	14
Signatix	2005-03-02	23	C/C++	C	2
Signatix	2005-03-02	28	Unix/Linux	Linux	10
Signatix	2005-03-02	39	Network	network	22
Signatix	2005-03-02	45	Network	network	22
Signatix	2005-03-02	46	Network	networking	22
Signatix	2005-03-02	49	C/C++	C	2
Signatix	2005-03-02	49	Database skills	Database	13
Signatix	2005-03-02	49	Network	network	22
Signatix	2005-03-02	49	C/C++	C	2
Signatix	2005-03-02	49	Database skills	Database	13
Signatix	2005-03-02	49	Network	network	22
Signatix	2005-03-02	52	Unix/Linux	Linux	10
Signatix	2005-03-02	52	Others OS	Solaris	12
Signatix	2005-03-02	53	Unix/Linux	Linux	10
Signatix	2005-03-02	54	Unix/Linux	Linux	10
Signatix	2005-03-02	55	Unix/Linux	Linux	10
Signatix	2005-03-02	65	Database skills	Database	13
Signatix	2005-03-02	65	PL/SQL	SQL	18
Signatix	2005-03-02	66	MS SQL Server	Microsoft S...	17
Signatix	2005-03-02	66	PL/SQL	SQL	18
Signatix	2005-03-02	66	Web skills	HTML	19
Signatix	2005-03-02	68	Java	Java	3
Signatix	2005-03-02	68	Oracle	Oracle	14
Signatix	2005-03-02	68	Java	Java	3

For example. A script might extract the text after “Name:” and put the text that follows into a column on a spreadsheet with the column header “Name”



Understanding Data Formats: Unstructured: Scraper Wiki

message Newsletter



PROFESSIONAL SERVICES HELP BLOG LOG IN

Liberate your data with ScraperWiki!

ScraperWiki helps you do data science on the web.

Get, clean, analyse, visualise and manage your data,
with simple tools or custom-written code.



ScraperWiki for businesses

Sign up – Free



I usually explore data in Excel, but
I want to take my data analysis to
the next level.



I want a place to write and run
code that does awesome stuff
with data.



I'm a business and I need
professional data extraction,
analysis and management.

Super-charge your data analysis with tools including...



Internews
Local voices. Global change.

Finding Data on the Web

- Understanding Data Formats
- **Advanced Google Searches**
- Scraping data from the web
- Exercise
- Summary & Resources

Finding Data on the Web:

Google Advanced Searches

Website Overview

Advanced Search

Find pages with...

all these words:

this exact word or phrase:

any of these words:

none of these words:

numbers ranging from:

to

To do this in the search box

Type the important words: `tricolor rat terrier`

Put exact words in quotes: `"rat terrier"`

Type OR between all the words you want: `miniature OR standard`

Put a minus sign just before words you don't want:
`-rodent, -"Jack Russell"`

Put 2 periods between the numbers and add a unit of measure:
`10..35 lb, $300..$500, 2010..2011`

Then narrow your results by...

language:

Find pages in the language you select.

region:

Find pages published in a particular region.

last update:

Find pages updated within the time you specify.

site or domain:

Search one site (like `wikipedia.org`) or limit your results to a domain like `.edu`, `.org` or `.gov`.

terms appearing:

Search for terms in the whole page, page title, or web address, or links to the page you're looking for.

SafeSearch:

Tell **SafeSearch** whether to filter sexually explicit content.

reading level:

Find pages at one reading level or just view the level info.

file type:

Find pages in the format you prefer.

File: www.google.com/advanced_search Tue May 20 2014 19:10:11 GMT-0700 (E. Africa Standard Time)



Internews
Local voices. Global change.

Finding Data on the Web

Two Options

Google Advanced Search:

http://www.google.com/advanced_search

Google Search with shortcuts

www.google.com

Finding Data: Advanced Google Searches Shortcuts

Type OR to find variations of the same search: Sri Lanka foreign assistance OR aid OR grant

Use quotes to search for phrases “Federation of University Teachers Association”

Narrow search to specific website: “site:url” Example site: site:http://www.moe.gov.lk/ higher education

Narrow search to filetype: “Filetype:[extension]”

Example: site:www.tradingeconomics.com filetype:pdf Sri Lanka

Google alerts: get notified about your beat

Finding Data on the Web

- **Understanding Data Formats**
- Exercise
- Advanced Google Searches
- **Scraping data from the web**
- Exercise
- Summary & Resources

Contents

Scraping data from the web

- **Using browser plug-ins to scrape data tables**
- Exercise
- Using Google docs to import data into charts
- Exercise

Summary & Resources



Internews
Local voices. Global change.

Scraping data using plug-ins

- Browser scrapers:
 - [Dafizilla Table2Clipboard](#) for Mozilla Firefox
 - [Scraper Extension](#) for Chrome
- Allow you to select tables on a website, including specific rows and columns
- Copy the data tables and either create a Google spreadsheet or paste into a blank Excel file



Contents

Scraping data from the web

- Using browser plug-ins to scrape data tables
- **Exercise**
- Using Google docs to import data into charts
- Exercise

Summary & Resources



Internews
Local voices. Global change.

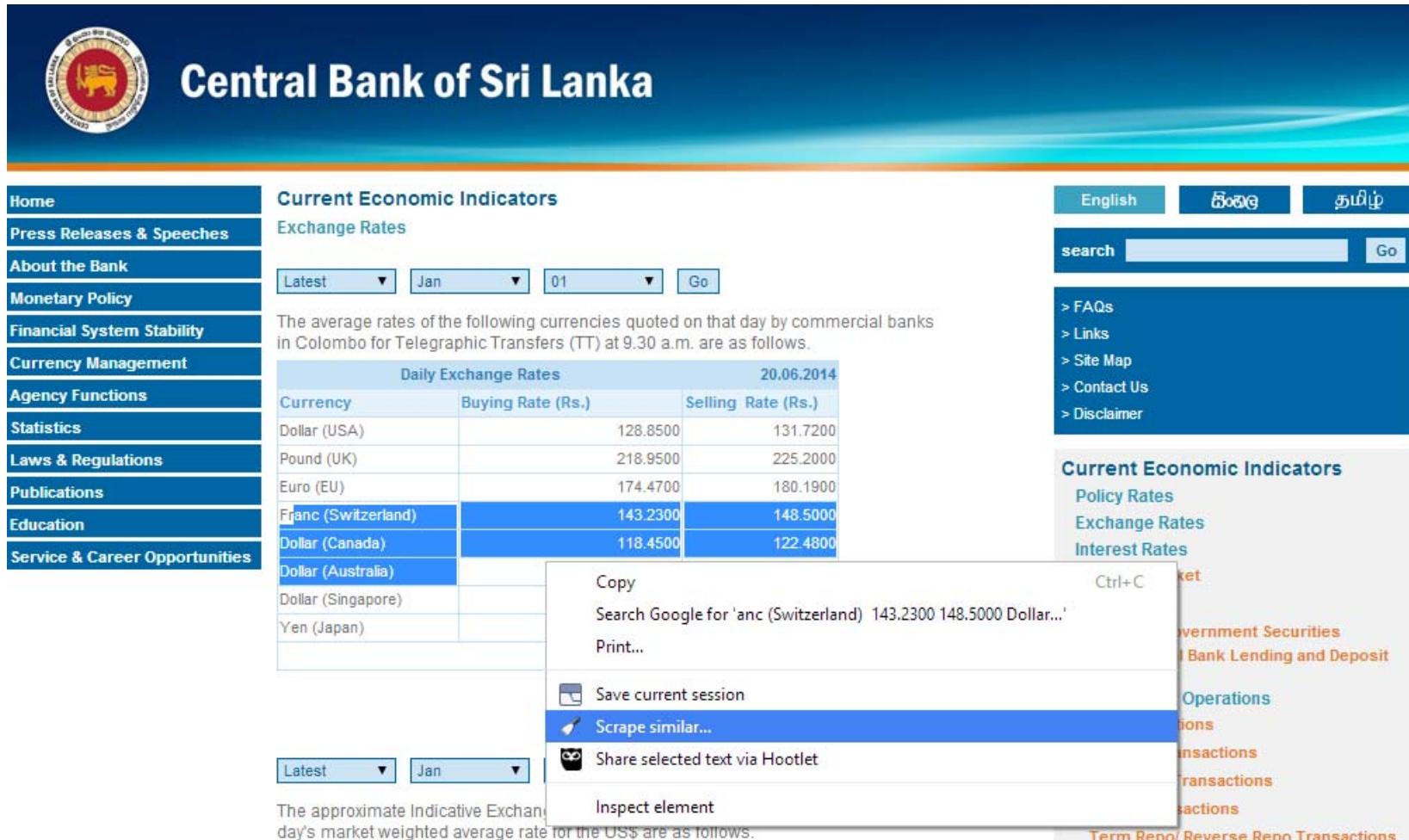
Scraping data using browser extensions exercise

1. The appropriate browser extension should be installed:
 - [Dafizilla Table2Clipboard](#) for Firefox
 - [Scraper Extension](#) for Chrome
2. Go to [Central Bank Current Economic Indicators](#).
3. If you are using the Chrome, highlight part of the table, right click and select 'Scrape similar' then Export to Google docs.

If you are using Mozilla, right click on the table and select "Table2Clipboard" and copy whole table. Then open an Excel file and paste.



Scraping data using browser extensions exercise



The screenshot shows the Central Bank of Sri Lanka website. The header includes the bank's logo and name. A navigation menu on the left lists various sections. The main content area displays 'Current Economic Indicators' with a sub-section for 'Exchange Rates'. It includes a table of daily exchange rates for various currencies as of 20.06.2014. A browser extension menu is open over the table, showing options like 'Copy', 'Search Google for...', 'Print...', 'Save current session', 'Scrape similar...', 'Share selected text via Hootlet', and 'Inspect element'. The 'Scrape similar...' option is highlighted.

Central Bank of Sri Lanka

Current Economic Indicators
Exchange Rates

Latest Jan 01 Go

The average rates of the following currencies quoted on that day by commercial banks in Colombo for Telegraphic Transfers (TT) at 9.30 a.m. are as follows.

Currency	Buying Rate (Rs.)	Selling Rate (Rs.)
Dollar (USA)	128.8500	131.7200
Pound (UK)	218.9500	225.2000
Euro (EU)	174.4700	180.1900
Franc (Switzerland)	143.2300	148.5000
Dollar (Canada)	118.4500	122.4800
Dollar (Australia)		
Dollar (Singapore)		
Yen (Japan)		

Latest Jan

The approximate Indicative Exchange Rates for the day's market weighted average rate for the US\$ are as follows.

Current Economic Indicators
Policy Rates
Exchange Rates
Interest Rates

English සිංහල தமிழ்

search Go

> FAQs
> Links
> Site Map
> Contact Us
> Disclaimer

Copy Ctrl+C
Search Google for 'anc (Switzerland) 143.2300 148.5000 Dollar...'
Print...
Save current session
Scrape similar...
Share selected text via Hootlet
Inspect element



Internews
Local voices. Global change.

Contents

Scraping data from the web

- Using browser plug-ins to scrape data tables
- Exercise
- **Using Google docs to import data into charts**
- Exercise

Summary & Resources

Scraping data using Google spreadsheets and Import HTML

1. Open
http://en.wikipedia.org/wiki/China_at_the_Olympics
2. Open a Google Spreadsheet
3. Type in
`=importHTML("http://en.wikipedia.org/wiki/China_at_the_Olympics","table",7)`
4. `=function("url", "object" , number)`



Scraping data using Google spreadsheets and Import HTML

1. Open
http://en.wikipedia.org/wiki/China_at_the_Olympics
2. Open a Google Spreadsheet
3. Type in
`=importHTML("http://en.wikipedia.org/wiki/China_at_the_Olympics","table",7)`
4. `=function("url", "object" , number)`



Scraping data using Google spreadsheets and Import HTML

Webpage Screenshot

Untitled spreadsheet

File Edit View Insert Format Data Tools Add-ons Help

Comments Share

Eva Constantaras

fx =importHTML("http://en.wikipedia.org/wiki/China_at_the_0lympics","table",7)

=importHTML("http://en.wikipedia.org/wiki/China_at_the_0lympics","table",7)

Sheet1

https://docs.google.com/spreadsheets/d/1cWUcQ0eeTTCYc-QsH_KekVdamdLbs-qskOgonw1XHDw/edit#gid=0 Fri May 23 2014 10:50:29 GMT+0300 (E. Africa Standard Time)

Contents

Scraping data from the web

- Using browser plug-ins to scrape data tables
- Exercise
- Using Google docs to import data into charts
- **Exercise**

Summary & Resources

Scraping data exercise

1. Visit [Trading Economics Sri Lanka Economic Indicators](#).
2. Find economic data that might enrich your reporting on the economy.
3. Try to scrape the data using both browser extensions and by importing from Google spreadsheets
4. Can you find a better way to search the site? Try “filetype:” and “site:” shortcuts that you learned in the session on searching for data on the web.



Finding Data on the Web: Exercise

1. Return to your lists of data that would enrich economic coverage
2. Search for data using databases and advanced searches.
3. Categorize the data you have by story angle

Finding Data on the Web: Exercise

Get back into your two groups:

- [Sri Lanka solving China's unemployment!](#)
Daily FT, June 19, 2013
- [Sri Lanka can be integral part of Asian transformation: ADB Chief](#), Lanka Business Online, June 19, 2014

Summary and Resources

Need help scraping your data from the web?

- Ask the School of Data
- Join the data driven journalism listserve
- Join the Hacks/Hackers Nairobi Google group and go to the next meeting or post your question and ask for volunteers
- Visit the Internews data journalism team for
- support
- Register for a free import.io account and try it out
- Try out [Outwit Hub](#)

Finding Data on the Web: Exercise

1. Return to your data angles
2. Use Advanced Searches to find related data
3. List data that you are missing to tell a complete story

Finding Data: No Excuses!

- “We don’t have that data on a computer.”
- High fees
- Delay tactics
- “Your request was unclear.”
- Sending the wrong data
- “Our database is too complicated to give you access.”
- “Our database software is proprietary.”
- “That information is protected by privacy law.”



Finding Data on the Web

- **Understanding Data Formats**
- Exercise
- Advanced Google Searches
- Scraping data from the web
- Exercise
- **Summary & Resources**

Summary & Resources

- Data is available in many different formats and almost always you will need to convert data to an Excel or csv file.
- Every day, more data is available online and accessible through searches
- If you can't find it online, don't give up!