FINDING DATA FOR STORIES

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

Internews
Local voices. Global change.

Where is Data?

Webpage Screenshot

STANDARD Digital

Last Updated: May 20, 2014 11:59

News | KTN | Radio Maisha | Business | The Nairobian | Sports | Entertainment | Evewoman | uReport

Your are here » Home » Kenya

10,000 missing files found at lands headquarters in Nairobi

By JAMES MBAKA

Updated Monday, May 12th 2014 at 22:32 GMT +3

Share this story:

Lands Cabinet Secretary Charity Ngilu at the central registry in Ardhi House, Nairobi, where auditors discovered files hidden by ministry staff. [PHOTO: BEVERLYNE MUSILI/STANDARD]

Glance Fact

Audit team stumbles upon stash of 'missing' documents stored by officers suspected to be involved in fraudulent transactions

Related Stories

Police kill seven suspected thugs in Nairobi

Ngilu promises efficiency as

FOLLOW US TODAY

Trending Now

1 **Avoid city, Sonu says on strike plans**
University of Nairobi student leaders have warned motorists to keep off major roads leadin...

2 **Mwanaisha Chidzuga: 'VIP treatment is not my thing'**
I have mastered the art of avoiding traffic and my husband has always been accorded that V...

3 **5 things couples do to avoid having sex**
There's no doubt that sex sells, but when it comes to sex in long-

http://www.standardmedia.co.ke/?articleID=2000119918&story_title=10-000-hidden-land-files-found-in-ministry Tue May 20 2014 12:29:59 GMT+0300 (E. Africa Standard Time)

The Standard, May 12, 2014

Internews
Local voices. Global change.

Contents

Contents

Internews
Local voices. Global change.

Understanding Data Formats

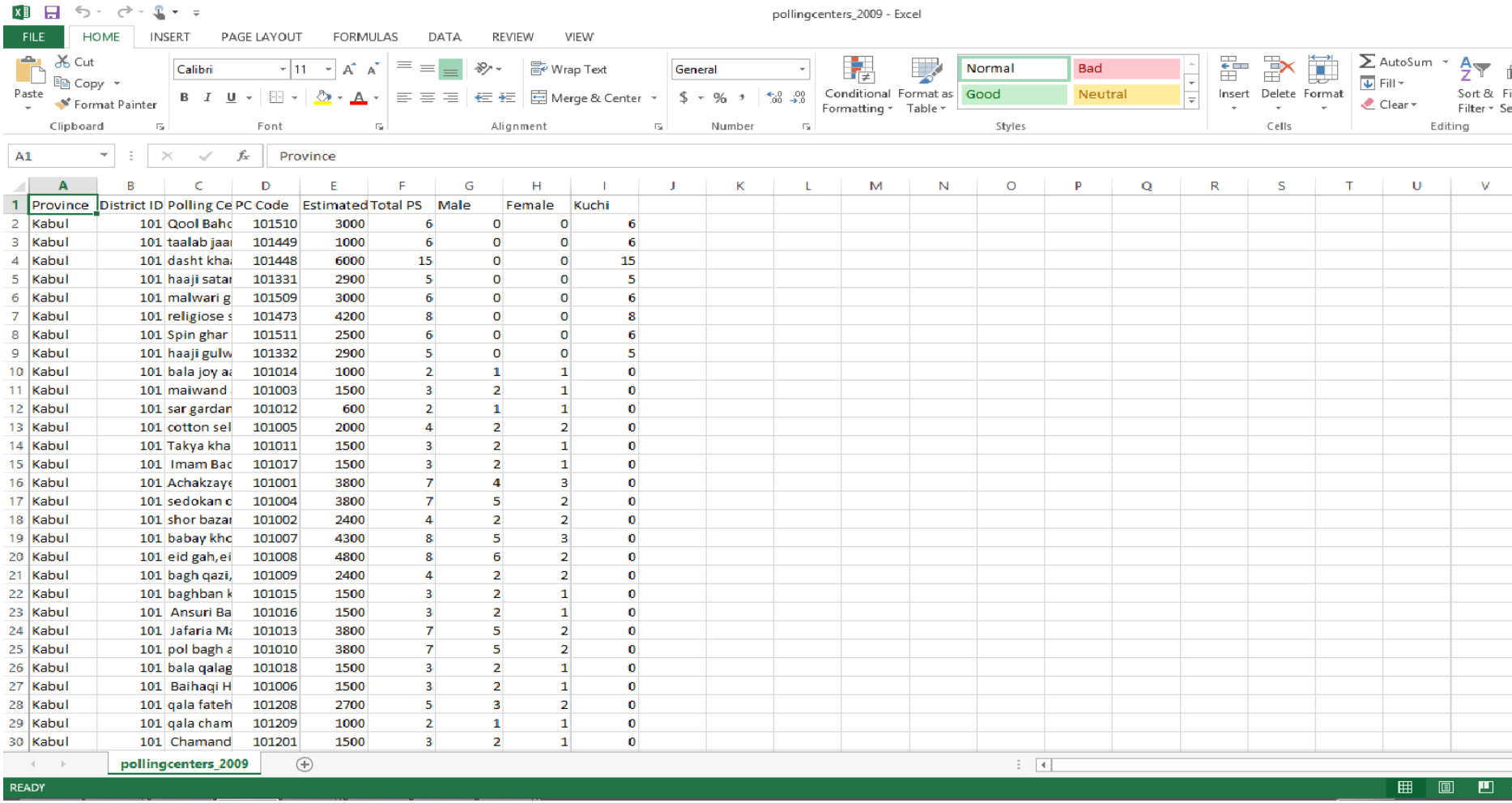
Machine readable

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

Unstructured

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

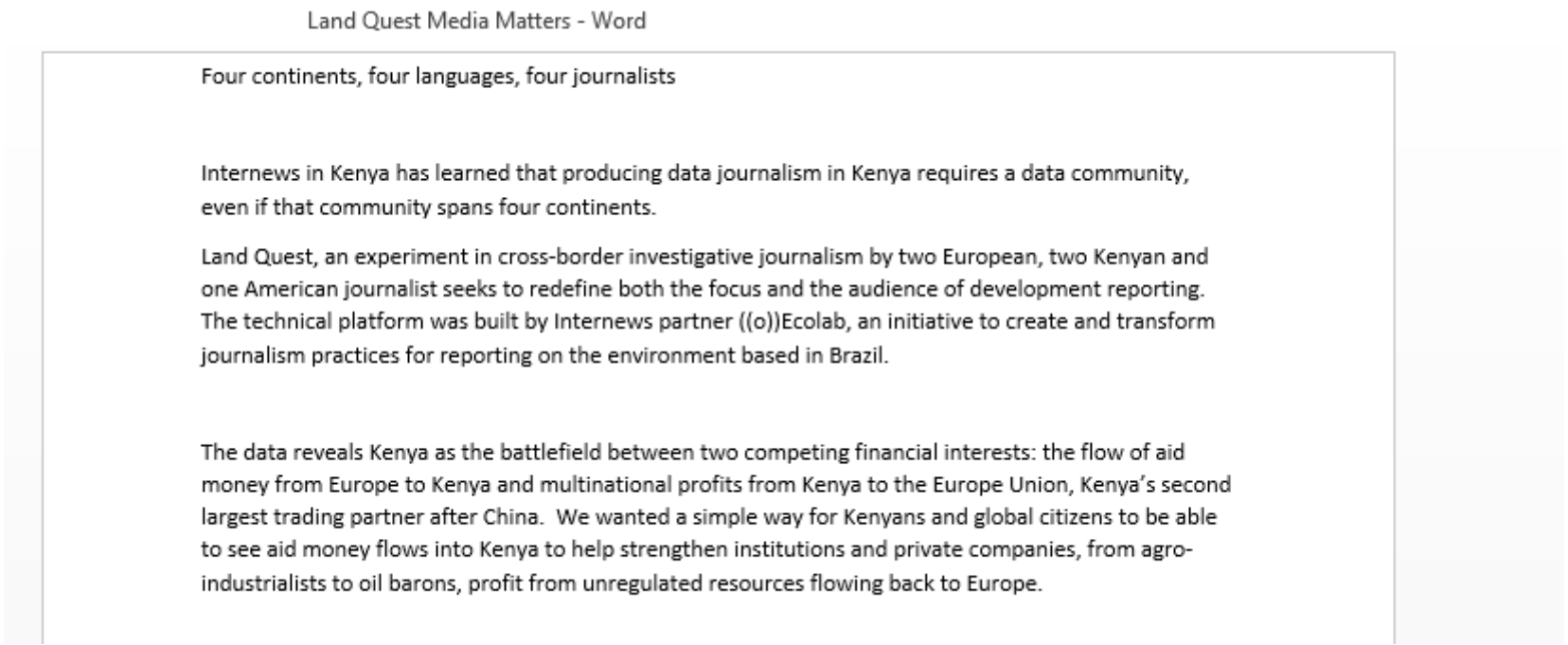
Understanding Data Formats: Machine Readable



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 jaa	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 gardar	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 bazar	101002	2400	4	2	2	0
Kabul	101	babay khc	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

Summary & Resources

Internews
Local voices. Global change.

Understanding Data Formats

PDFs

Scanned Images (Unstructured, not searchable)

Computer-generated (searchable)

Tables (Structured, searchable)

Complex formats (Unstructured, searchable)

Internews
Local voices. Global change.

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

Internews
Local voices. Global change.

Understanding Data Formats: Exercise

1. Open the File: **Country Profile President's Malaria Initiative (PMI)** or **Maternal Index**
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

   pdftoexcelonline.com

   zamzar.com

   pdftoexcel.org

Internews
Local voices. Global change.

Contents

Internews
Local voices. Global change.

Understanding Data Formats: Scanned Images

Optical Character Recognition Software

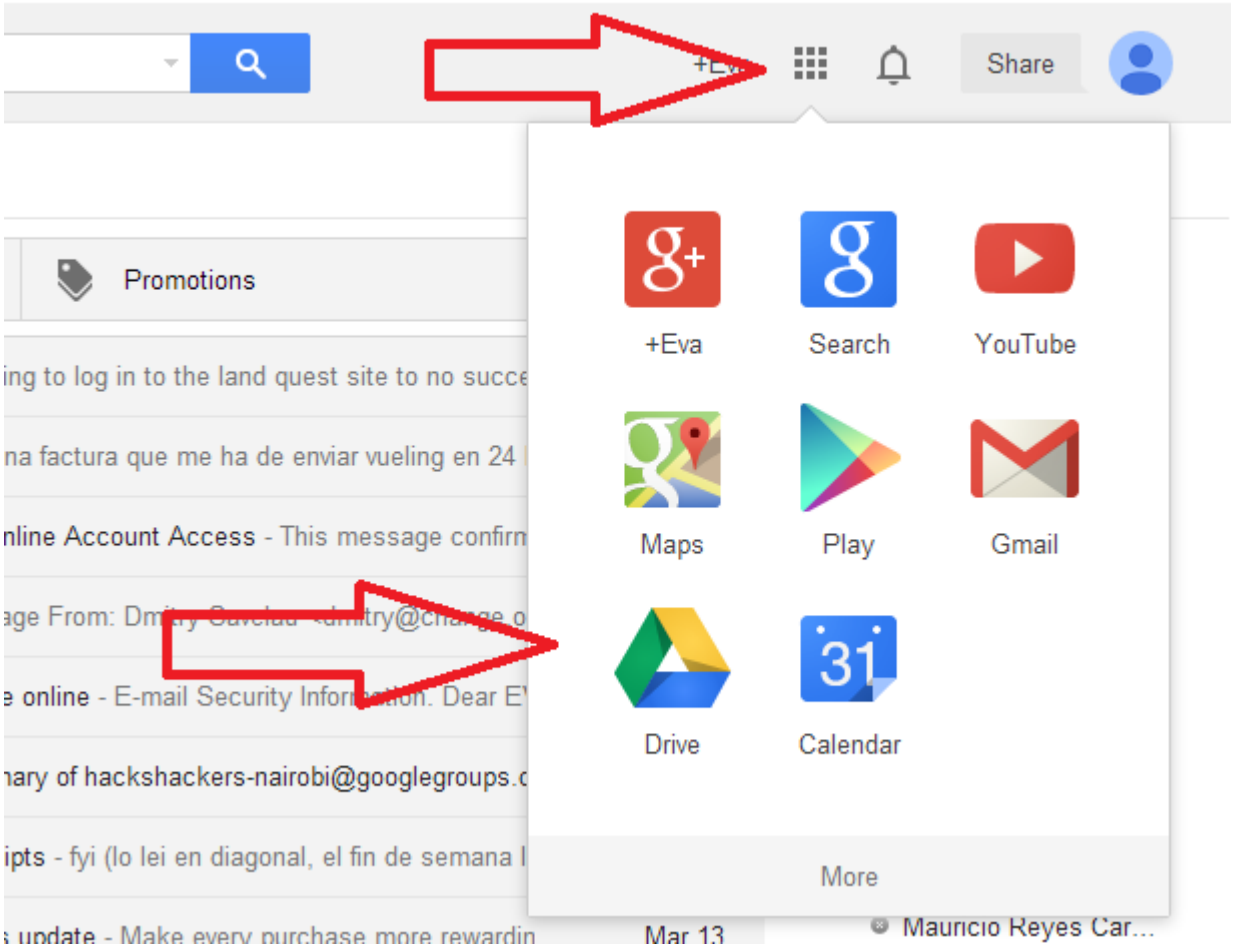
- For documents that you can't search or select text
- Quality depends on quality of image, clarity of text
- Adobe Acrobat Professional: paid
- Google Docs: free
- Document Cloud: free. Advantage is that it allows you to annotate and publish your document as part of a multimedia publication. Your media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

Contents

Internews
Local voices. Global change.

Understanding Data Formats: Scanned Images Exercise

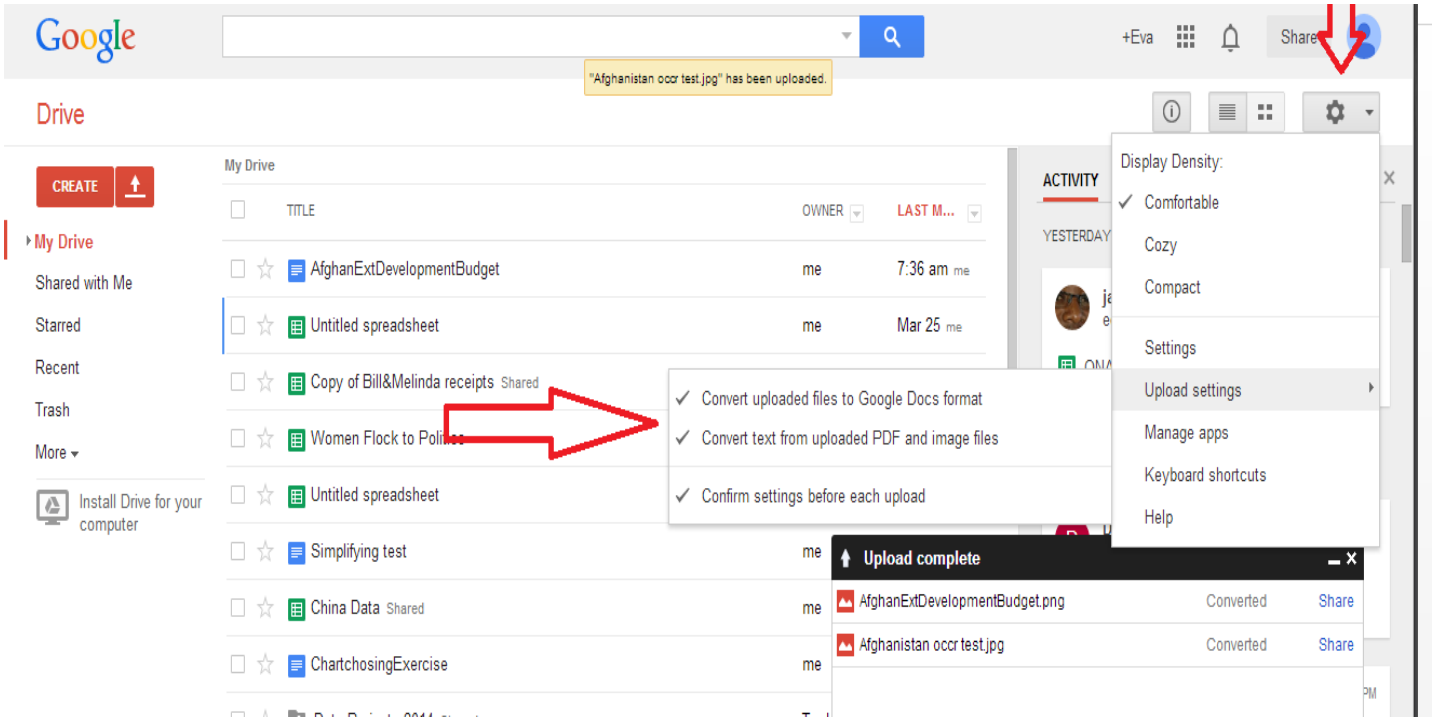
1. Open **Specimen Financial Statement**
2. Open Google Drive (click on grid in the top right corner of G-mail and select Drive)



Internews
Local voices. Global change.

Understanding Data Formats: Scanned Images Exercise

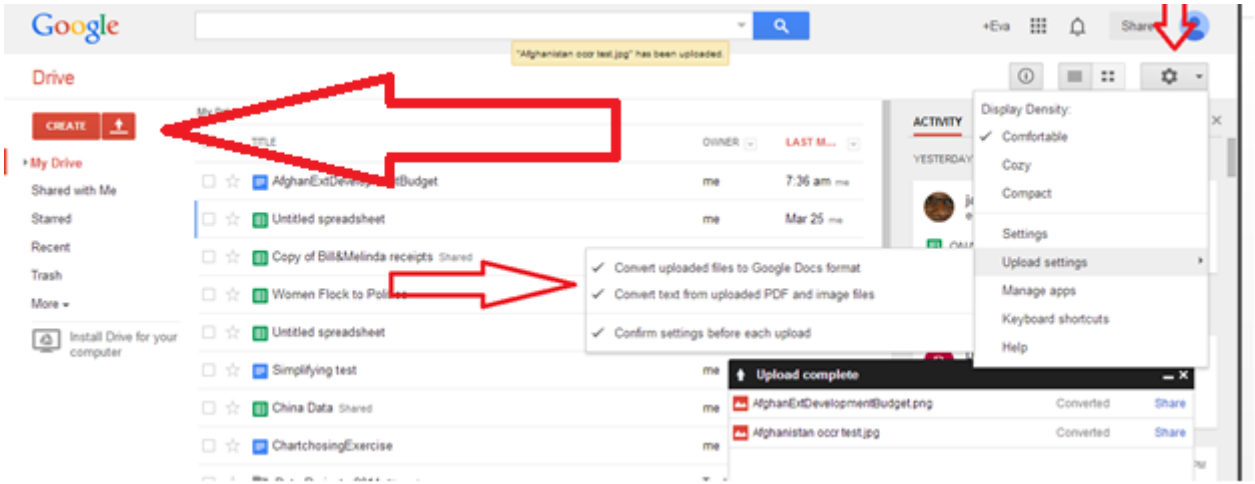
3. Go to Setting and Upload Settings: Select all options



Internews
Local voices. Global change.

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.



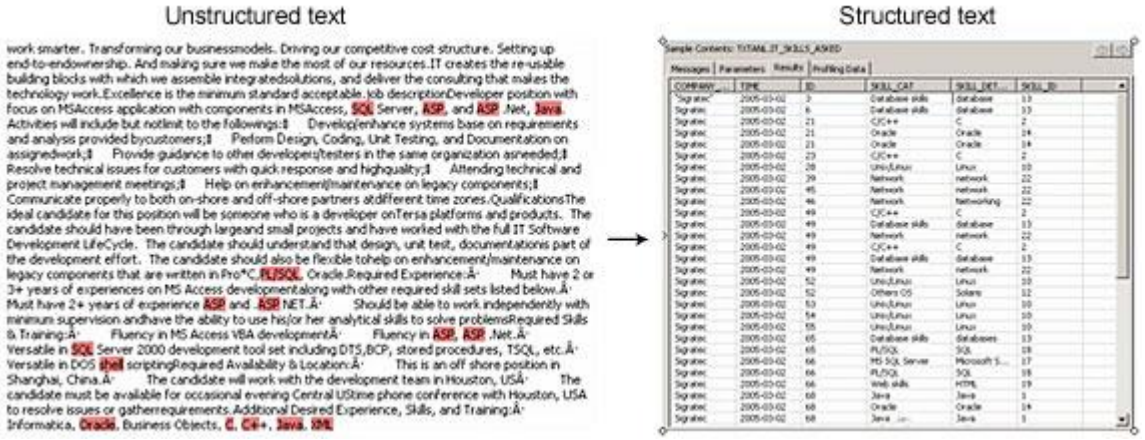
Internews
Local voices. Global change.

Contents

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



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”

Internews
Local voices. Global change.

Understanding Data Formats: Unstructured: Scraper Wiki

Webpage Screenshot

Contents

Summary and Resources

Need help liberating your data?

- Download Tabula tabula.nerdpower.org
- 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
- Join our session for data scraping without programming

Internews
Local voices. Global change.