

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

Where is Data?



News

KTN

Radio Maisha

Business

The Nairobiian

Sports

Entertainment

EveWoman

uReport



You 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-

[The Standard](#), May 12, 2014



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

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)

Understanding Data Formats: Machine Readable

pollingcenters_2009 - Excel

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

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

Understanding Data Formats

PDFs

Scanned Images
(Unstructured, not
searchable)

Computer-generated
(searchable)

Tables (Structured,
searchable)

Complex formats
(Unstructured,
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

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: 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
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

Summary & Resources

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

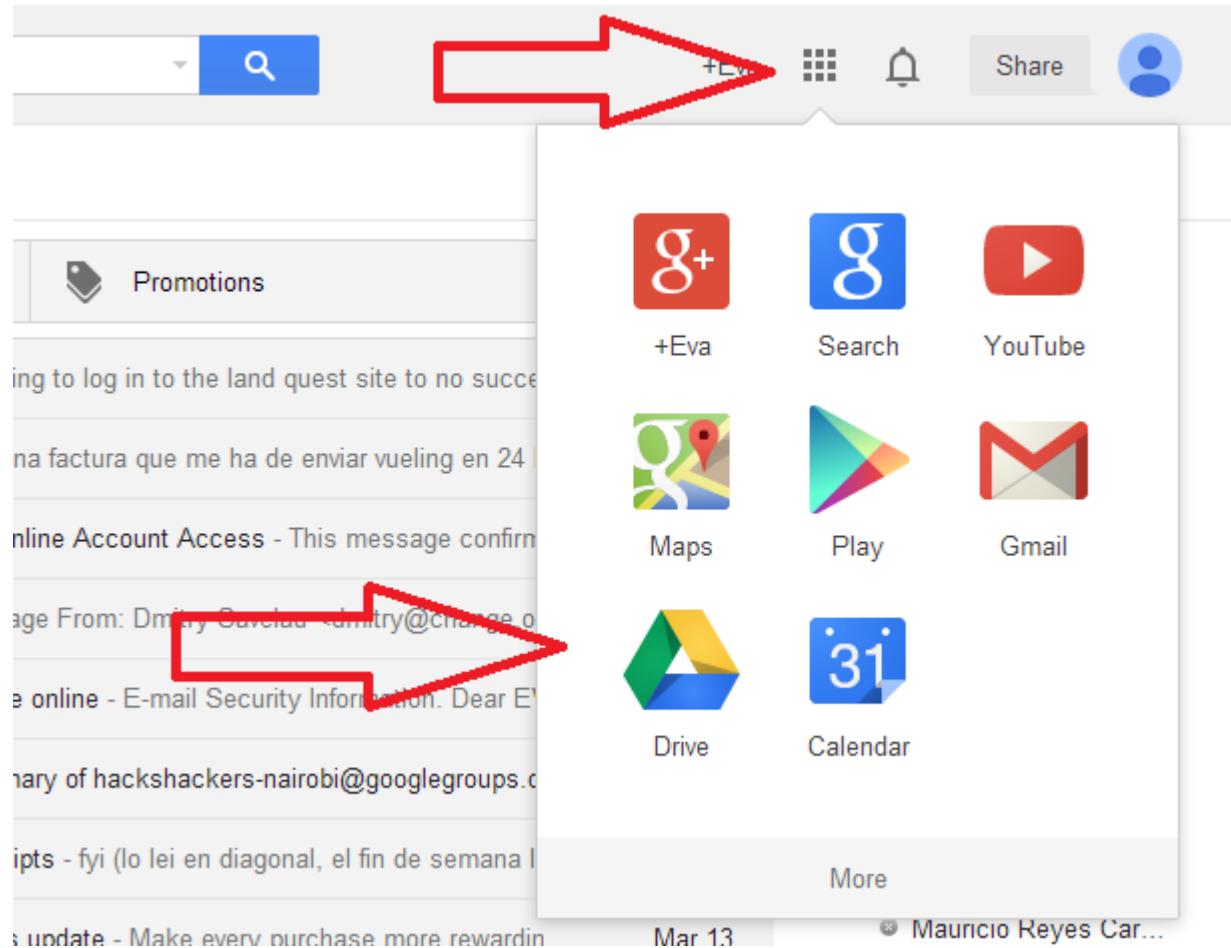
Understanding Data Formats

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

Summary & Resources

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)



Understanding Data Formats: Scanned Images Exercise

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

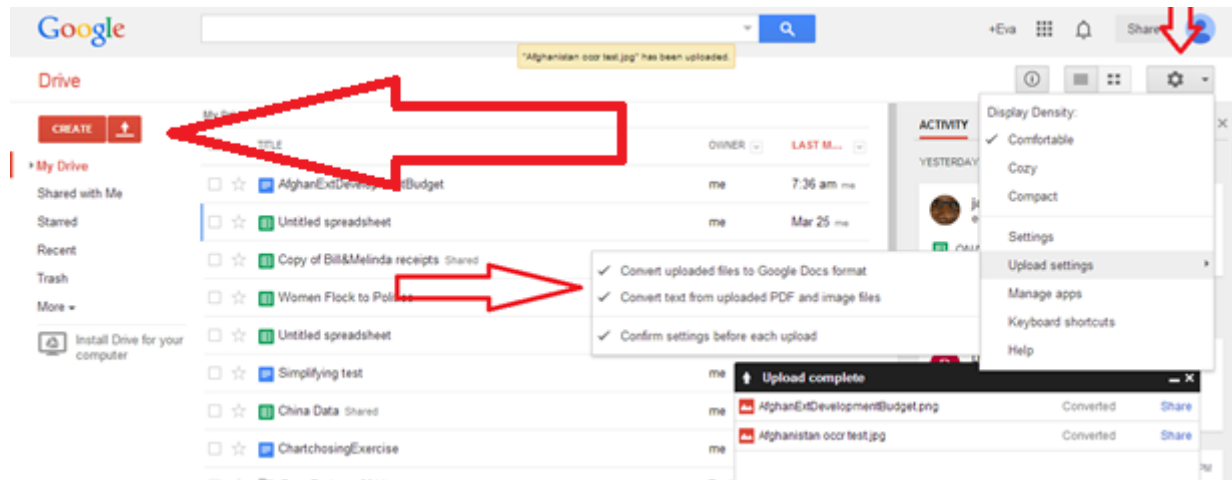
The screenshot shows the Google Drive web interface. At the top, a notification bar states: "Afghanistan occr test.jpg" has been uploaded. Below this, the "My Drive" section lists several files. A red arrow points to the "Upload settings" option in the "Settings" menu. The "Upload settings" menu is open, showing three checked options: "Convert uploaded files to Google Docs format", "Convert text from uploaded PDF and image files", and "Confirm settings before each upload". A red arrow also points to the "Share" button in the top right corner. At the bottom, an "Upload complete" notification shows two files: "AfghanExtDevelopmentBudget.png" and "Afghanistan occr test.jpg", both marked as "Converted".

TITLE	OWNER	LAST M...
AfghanExtDevelopmentBudget	me	7:36 am me
Untitled spreadsheet	me	Mar 25 me
Copy of Bill&Melinda receipts Shared		
Women Flock to Polit...		
Untitled spreadsheet		
Simplifying test	me	
China Data Shared	me	
ChartchoosingExercise	me	

Upload complete		
AfghanExtDevelopmentBudget.png	Converted	Share
Afghanistan occr test.jpg	Converted	Share

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**

Summary & Resources

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 developers/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 Terza 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 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 batch 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

Sample Contents: TITM12_SKILLS_ASKED					
Messages	Parameters	Results	Profiling Data		
COMPANY	TIME	ID	SKILL_CAT	SKILL_DET	SKILL_ID
Signature	2005-03-02	2	Database skills	Database	13
Signature	2005-03-02	6	Database skills	Database	13
Signature	2005-03-02	21	C/C++	C	2
Signature	2005-03-02	21	Oracle	Oracle	14
Signature	2005-03-02	21	Oracle	Oracle	14
Signature	2005-03-02	23	C/C++	C	2
Signature	2005-03-02	28	Unix/Linux	Linux	10
Signature	2005-03-02	29	Network	network	22
Signature	2005-03-02	45	Network	network	22
Signature	2005-03-02	46	Network	networking	22
Signature	2005-03-02	49	C/C++	C	2
Signature	2005-03-02	49	Database skills	Database	13
Signature	2005-03-02	49	Network	network	22
Signature	2005-03-02	49	C/C++	C	2
Signature	2005-03-02	49	Database skills	Database	13
Signature	2005-03-02	49	Network	network	22
Signature	2005-03-02	52	Unix/Linux	Linux	10
Signature	2005-03-02	52	Others OS	Solaris	12
Signature	2005-03-02	53	Unix/Linux	Linux	10
Signature	2005-03-02	54	Unix/Linux	Linux	10
Signature	2005-03-02	55	Unix/Linux	Linux	10
Signature	2005-03-02	65	Database skills	Database	13
Signature	2005-03-02	65	PL/SQL	SQL	18
Signature	2005-03-02	66	MS SQL Server	Microsoft S...	17
Signature	2005-03-02	66	PL/SQL	SQL	18
Signature	2005-03-02	66	Web skills	HTML	19
Signature	2005-03-02	68	Java	Java	5
Signature	2005-03-02	68	Oracle	Oracle	14
Signature	2005-03-02	68	Java	Java	5

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

Webpage Screenshot



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...

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

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