Báo chí dữ liệu môi trường
Nhu cầu mở cho Việt Nam và khu vực Mê Kông

HÀ NỘI, 26-28/10/2016

Trích xuất dữ liệu từ web

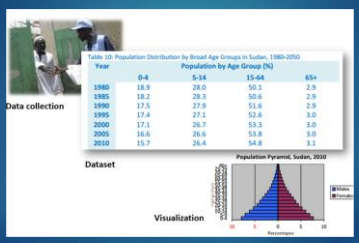
CÁC PHƯƠNG PHÁP VÀ CÔNG CỤ TRÍCH XUẤT DỮ LIỆU ONLINE

Trích xuất dữ liệu từ Web

- **Hiểu định dạng dữ liệu**

Sử dụng các tiện ích của trình duyệt để trích xuất bảng số liệu
Dùng Google docs để nhập dữ liệu vào biểu đồ
Chuyển file PDF thành Excel dùng Tabula
Chuyển file PDF thành Excel dùng công cụ online
Tóm tắt

Vòng đời của dữ liệu



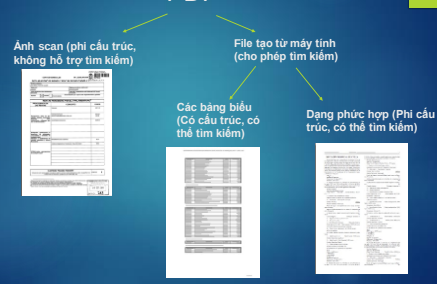
Dữ liệu lưu trữ dạng truyền thống…

Dữ liệu lưu trữ kiểu truyền thống…

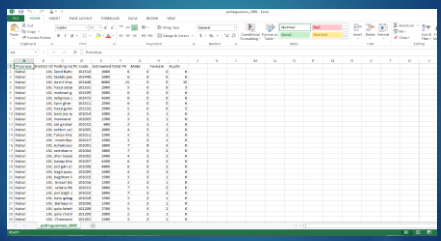
Các định dạng dữ liệu tải về

- Định dạng tài liệu di động (PDF): hỗ trợ biểu đồ đi kèm với text một cách đồng bộ
- File Excel (XLS): Dữ liệu bảng biểu cho phép máy tính đọc được, dùng phần mềm Microsoft Excel
- Dạng tập tin CSV: Định dạng văn bản thuần, sử dụng dấu phẩy để tách biệt dữ liệu (dùng để chuyển dữ liệu/bảng tính giữa các ứng dụng khác nhau

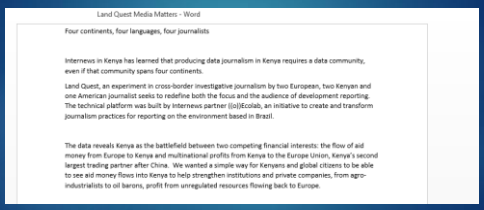
PDF



Dữ liệu máy tính có thể đọc



Dữ liệu phi cấu trúc


Land Quest Media Matters - Word

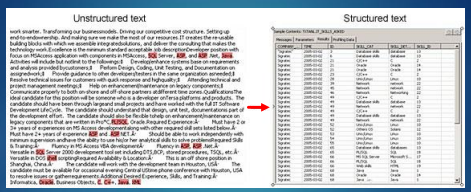
Four continents, four languages, four journalists

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.

Dữ liệu phi cấu trúc



Chẳng hạn, kịch bản này cho phép nhận dạng và tách phần dữ liệu với "tiêu đề" để tạo thành bảng dữ liệu với các tiêu đề cột được giữ nguyên.

Dữ liệu phi cấu trúc: Công cụ Scraper Wiki

Trích xuất dữ liệu từ Web

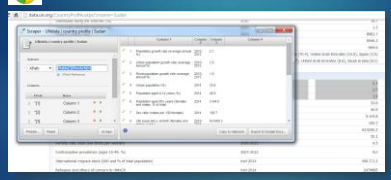
Hiểu định dạng dữ liệu
- **Sử dụng các tiện ích hỗ trợ trình duyệt để trích xuất bảng số liệu**

Dùng Google docs để nhập dữ liệu vào biểu đồ
Chuyển file PDF thành Excel dùng Tabula
Chuyển file PDF thành Excel dùng công cụ online
Tóm tắt

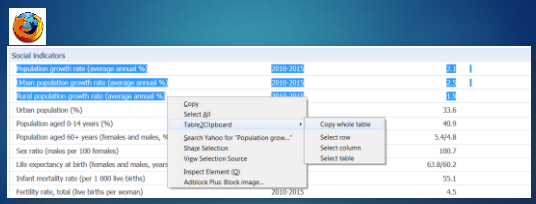
Dùng tiện ích của trình duyệt

- Các tiện ích trích xuất dữ liệu từ trình duyệtBrowser scrapers:
  - Trình duyệt Mozilla Firefox: Dafizilla Table2Clipboard
    https://addons.mozilla.org/en-us/firefox/addon/dafizilla-table2clipboard/
  - Trình duyệt Chrome: Scraper Extension
    https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd
- Cho phép chọn bảng trên webiste, gồm cả hàng và cột
- Copy bảng dữ liệu và tạo bảng tính với Google Spreadsheet hoặc dán file Excel.

Using Browser Plug-Ins



Dùng tiện ích của trình duyệt

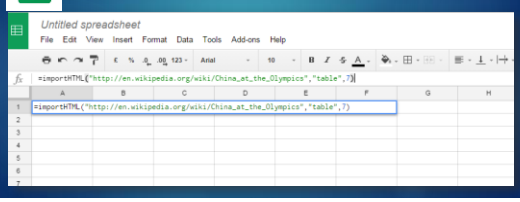


Trích xuất dữ liệu từ Web

Hiểu định dạng dữ liệu
Sử dụng các tiện ích hỗ trợ trình duyệt để trích xuất bảng số liệu
- **Dùng Google docs để nhập dữ liệu vào biểu đồ**

Chuyển file PDF thành Excel dùng Tabula
Chuyển file PDF thành Excel dùng công cụ online
Tóm tắt

Trích xuất dữ liệu dùng bảng Google spreadsheets và câu lệnh Import HTML



Trích xuất dữ liệu dùng bảng Google spreadsheets và câu lệnh Import HTML

1. Mở http://en.wikipedia.org/wiki/China_at_the_Olympics
2. Mở file Google Spreadsheet
3. Gõ câu lệnh =importHTML("http://en.wikipedia.org/wiki/China_at_the_Olympics","table",7)
4. =function("url", "object" , number)

Trích xuất dữ liệu từ Web

Hiểu định dạng dữ liệu
Sử dụng các tiện ích hỗ trợ trình duyệt để trích xuất bảng số liệu
Dùng Google docs để nhập dữ liệu vào biểu đồ
- **Chuyển file PDF thành Excel dùng Tabula**

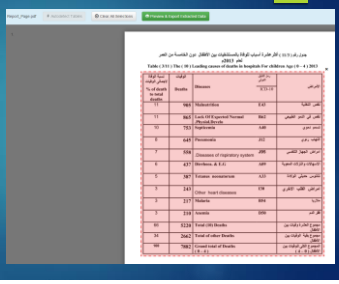
Chuyển file PDF thành Excel dùng công cụ online
Tóm tắt

Công cụ chuyển đổi từ file PDF sang Excel offline

Công cụ offline chuyển đổi từ file PDF sang Excel

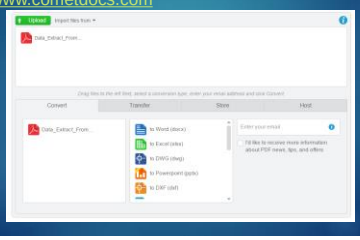
Thao tác với Tabula

3. Chọn bảng dữ liệu bằng cách click chuột vào góc trên trái và rê đến hết góc dưới phải của bảng. Nếu chọn chuẩn, cả bảng được chọn phải chuyển màu



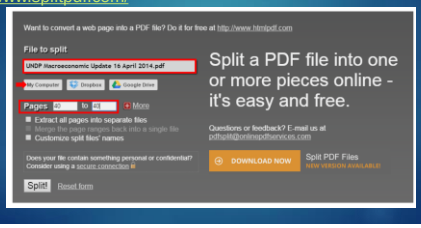
Công cụ online chuyển đổi từ file PDF sang Excel

www.cometdocs.com



Công cụ online chuyển đổi từ file PDF sang Excel

www.splitpdf.com/



Trích xuất dữ liệu từ Web

Hiểu định dạng dữ liệu
Sử dụng các tiện ích hỗ trợ trình duyệt để trích xuất bảng số liệu
Dùng Google docs để nhập dữ liệu vào biểu đồ
Chuyển file PDF thành Excel dùng Tabula
Chuyển file PDF thành Excel dùng công cụ online
- **Tóm tắt**

Cách làm sạch dữ liệu đơn giản

- Xác định nguồn
- Chuẩn bị (prepping) bộ dữ liệu
- Chỉnh sửa nhãn
- Tìm kiếm và thay thế
- Chuyển đổi định dạng
- Sắp xếp dữ liệu

Làm sạch dữ liệu là gì?

- Tìm kiếm và loại bỏ những phần dữ liệu không mong muốn trong bảng tính
- Định dạng lại dữ liệu để có thể sử dụng các công cụ tính toán
- Điều chỉnh sự thiếu nhất quán của dữ liệu
- Cấu trúc lại dữ liệu để có thể khai thác hiệu quả, đúng với nhu cầu nhất

Tại sao phải làm sạch dữ liệu?

- Giảm thiểu ảnh hưởng của việc thiếu hụt dữ liệu
- Sửa lỗi trong dữ liệu
- Loại bỏ các dữ liệu trùng lặp
- Loại bỏ các thông tin không mong muốn

Tóm tắt

- Dữ liệu thường bị "trói" ở định dạng PDFs, dưới dạng có cấu trúc hoặc phi cấu trúc, có hỗ trợ công cụ tìm kiếm hoặc không cho phép tìm kiếm.
- Việc nhận biết định dạng của file PDF rất quan trọng trước khi chọn công cụ chuyển đổi.
- Hiện đã có các phần mềm online cho phép chuyển đổi bảng dữ liệu ở định dạng PDF cho phép tìm kiếm sang bảng dữ liệu Excel hoặc CSV
- Trích xuất trực tiếp dữ liệu từ web là một giải pháp trong trường hợp trang web không cho phép tải dữ liệu hoặc công cụ tải dữ liệu gặp trục trặc.
- Tiện ích mở rộng của Trình duyệt web cho phép trích xuất dữ liệu trực tiếp từ web.
- Dùng chức năng Import HTML để kéo dữ liệu từ website về bảng tính Google spreadsheet.
- Với các dạng dữ liệu phức tạp hơn, có lẫn dữ liệu dạng text trong bảng số liệu, cần dùng ngôn ngữ lập trình để xử lý.

CẢM ƠN!