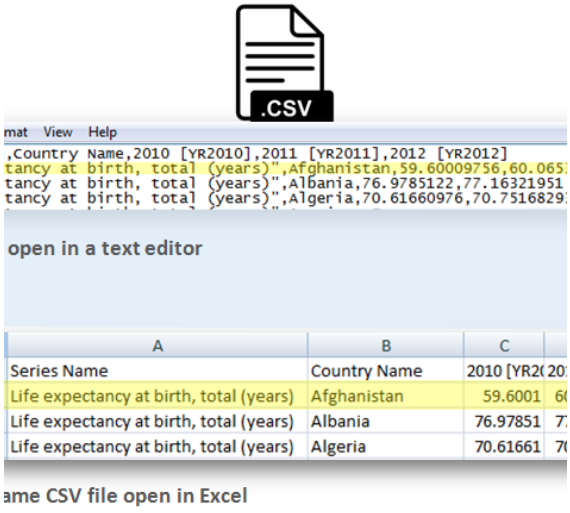
Các định dạng dữ liệu thường gặp

Hiểu về dữ liệu do máy tính tạo ra

Định dạng dữ liệu - Máy tính có thể đọc được, do máy tính tạo ra, có cấu trúc



Định dạng dữ liệu - Máy tính có thể đọc được, do máy tính tạo ra, có cấu trúc

- Với những định dạng này, phần mềm máy tính nhận dạng được cấu trúc rõ ràng của dữ liệu – phổ biến nhất là dạng bảng – với cột và dòng được sắp xếp và mô tả các điểm dữ liệu riêng biệt. VD: Excel và CSV.
- ***Tệp Excel (XLS, XLSX.):*** dữ liệu được lưu trữ ở dạng bảng có thể đọc được bằng phần mềm Microsoft Excel
- ***Comma separated values (CSV):*** Tệp văn bản đơn giản, các dữ liệu nhập vào được phân tách bằng dấu phẩy

Định dạng dữ liệu - Máy tính có thể đọc được, do máy tính tạo ra, có cấu trúc

- **Lưu ý:** định dạng CSV và TSV (tab-separated values) là định dạng để "mã hóa" dữ liệu dạng bảng.
- Các tệp CSV và TSV là các file text mà ở đó:
  - Mỗi dòng thể hiện một hàng và
  - Trong mỗi dòng, các cột được phân tách với nhau bằng dấu phẩy (đối với CSV) hoặc ký tự Tab (đối với TSV)
- Các tệp Excel cũng dùng cấu trúc tương tự, nhưng dựa trên phần mềm của Microsoft.

Ví dụ về dữ liệu bảng có cấu trúc, máy tính có thể đọc được

File Excel: Hiện trạng rừng có đến 31/12 phân theo địa phương

Nguồn: Tổng cục Thống kê

Hiện trạng rừng có đến 31/12 phân theo địa phương

Nghìn ha

	2008[*1]					2009				2010				
	Tổng diện tích rừng	Chia ra			Tỷ lệ che phủ rừng (%)	Tổng diện tích rừng	Chia ra		Tỷ lệ che phủ rừng (%)	Tổng diện tích rừng	Chia ra			Tỷ lệ che phủ rừng (%)
		Rừng tự nhiên	Rừng trồng				Rừng tự nhiên	Rừng trồng			Rừng tự nhiên	Rừng trồng		
			Tổng số	Mới trồng								Tổng số	Mới trồng(**)	
CẢ NƯỚC	**13118.8**	**10348.6**	**2770.2**	**342.7**	**38.7**	**13258.7**	**10338.9**	**2919.8**	**39.1**	**13388.1**	**10304.8**	**3083.3**	**357.1**	**39.5**
Đồng bằng sông Hồng	***416.4***	***212.8***	***203.6***	***36.7***		***428.9***	***207.6***	***221.3***		***434.9***	***203.4***	***231.5***	***30.4***	
Hà Nội	23.0	5.0	18.0	0.8	6.6	24.5	6.9	17.6	7.1	24.3	6.9	17.4	1.0	7
Vĩnh Phúc	28.4	9.4	19.0	1.5	21.8	28.6	9.4	19.2	22.3	28.5	9.4	19.2	1.0	22.4
Bắc Ninh	0.6		0.6	0.2	0.5	0.6		0.6	0.7	0.6		0.6		0.7
Quảng Ninh	291.3	155.9	135.4	32.2	42.6	301.8	149.2	152.6	44.4	310.4	147.3	163.0	26.8	46.2
Hải Dương	10.4	2.3	8.1	0.0	6.3	10.3	2.3	8.0	6.2	10.2	2.3	7.9		6.2
Hải Phòng	17.3	10.8	6.5	0.2	11.2	17.8	10.8	7.0	11.2	18.0	10.8	7.2	0.8	11.3
Thái Bình	7.5	0.0	7.5	0.7	4.4	7.7		7.7	4.9	7.3		7.3		4.8
Hà Nam	8.0	5.9	2.1	0.1	9.3	7.4	5.4	2.0	8.6	4.8	3.1	1.6	0.3	5.3
Nam Định	2.8	0.0	**2.8**	0.0	1.7	2.8		**2.8**	1.7	3.6		**3.6**		2.2
Ninh Bình	27.1	23.5	3.6	1.0	18.8	27.4	23.6	3.8	19.1	27.2	23.6	3.6	0.5	19.3
Trung du và miền núi ph	***4558.4***	***3574.5***	***983.9***	***123.2***		***4633.5***	***3565.8***	***1067.6***		***4675.0***	***3584.7***	***1090.3***	***145.0***	
Hà Giang	422.4	363.9	58.5	6.0	52.6	427.5	360.2	67.3	51.6	444.9	367.7	77.2	20.9	53.3
Cao Bằng	333.5	316.8	16.7	0.9	49.5	334.9	318.0	16.9	49.8	336.8	319.7	17.1	0.5	50
Bắc Kạn	274.3	228.7	45.6	3.2	55.7	281.3	230.0	51.3	56.6	288.1	229.0	59.1	8.9	57.5
Tuyên Quang	386.1	284.7	101.4	19.3	62.5	386.1	273.8	112.3	62.8	390.1	270.6	119.5	13.0	64.1
Lào Cai	314.9	253.3	61.6	9.6	47.8	323.3	257.7	65.6	49.4	327.8	258.4	69.3	8.1	50.1
Yên Bái	400.2	231.9	168.3	12.0	56.3	404.4	231.6	172.8	56.9	410.7	234.7	176.0	12.9	57.7
Thái Nguyên	167.9	99.9	68.0	7.6	45.3	171.7	98.6	73.1	45.7	175.1	97.0	78.1	11.4	46
Lạng Sơn	382.4	242.6	139.8	16.2	44.1	393.9	244.0	149.9	45.1	409.4	251.4	158.0	21.3	46.4

Ví dụ về dữ liệu bảng có cấu trúc, máy tính có thể đọc được

File CSV: Chỉ số phát triển rừng 2005-2017.

Nguồn: ODV

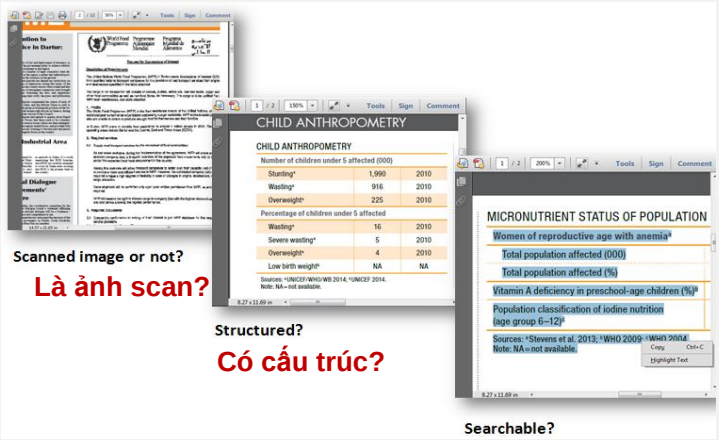
Nam	Tong_so	Rung_san_xuat	Rung_phong_ho	Rung_dac_dung
2005	96.1	96.9	92.8	90
2006	108.7	109.3	105.2	111.1
2007	98.5	97.3	105.3	105
2008	105.4	100.9	133.1	47.6
2009	121.4	122.5	114.6	220
2010	103.9	97.6	126.1	200
2011	84	101.9	26.3	59.1
2012	88.2	88	96.7	53.8
2013	121.4	123.9	96.6	85.7
2014	97.6	93.8	154.6	108.3
2015	112.8	113.5	106.9	100
2016	96	97.1	85.5	92.3
2017	100.5	101.3	91.9	100

Các công cụ

			
Excel hoặc Google sheets	Access hoặc SQL (quản lý dữ liệu)	SPSS hoặc STATA (phân tích thống kê)	Python hoặc R (lập trình để phân tích và trực quan hóa dữ liệu)

PDFS

Định dạng PDF - Các câu hỏi chính



Scanned image or not?

Là ảnh scan?

Structured?

Có cấu trúc?

Searchable?

Có thể tìm kiếm?

Định dạng PDF - Các câu hỏi chính

1. Là ảnh scan?

File này được máy tính xuất ra ở định dạng pdf hay là hình ảnh scan lại một văn bản đã được in ra?

2. Có cấu trúc?

Dữ liệu trong file PDF có được sắp xếp thành các dòng và cột theo bảng không?

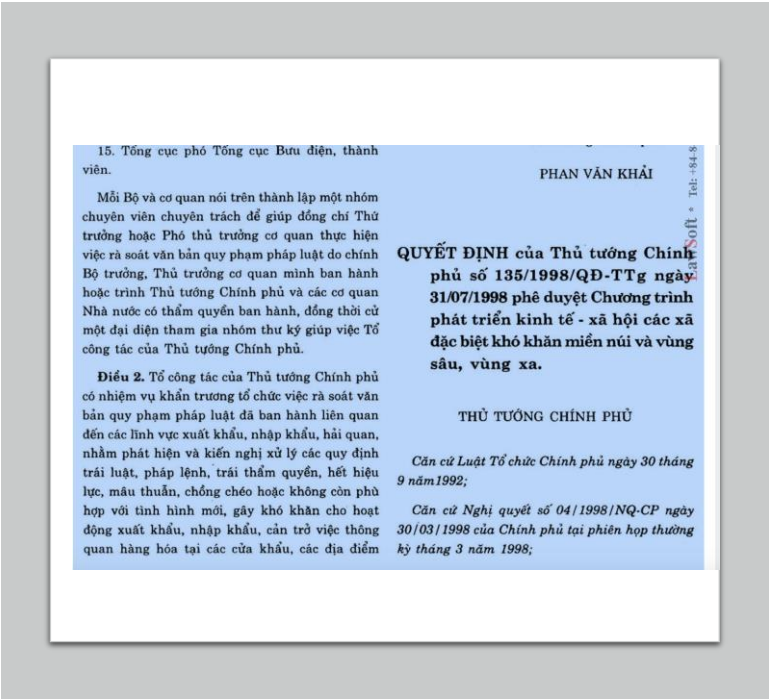
3. Có thể tìm kiếm trong văn bản không?

Với các file pdf được tạo ra từ máy tính, bạn có thể đánh dấu (highlight), và máy tính có thể nhận diện được chữ cái và số dưới dạng “ký tự”.

Ví dụ về một file PDF là ảnh scan

- Quyết định 135/1998/QĐ-TTG phê duyệt Chương trình Phát triển KT-XH các xã đặc biệt khó khăn miền núi và vùng sâu vùng xa

Nguồn: Cổng thông tin Điện tử Chính phủ.



15. Tổng cục phó Tổng cục Bưu điện, thành viên.

Mỗi Bộ và cơ quan nói trên thành lập một nhóm chuyên viên chuyên trách để giúp đồng chí Thứ trưởng hoặc Phó thủ trưởng cơ quan thực hiện việc rà soát văn bản quy phạm pháp luật do chính Bộ trưởng, Thủ trưởng cơ quan mình ban hành hoặc trình Thủ tướng Chính phủ và các cơ quan Nhà nước có thẩm quyền ban hành, đồng thời cử một đại diện tham gia nhóm thư ký giúp việc Tổ công tác của Thủ tướng Chính phủ.

Điều 2. Tổ công tác của Thủ tướng Chính phủ có nhiệm vụ khẩn trương tổ chức việc rà soát văn bản quy phạm pháp luật đã ban hành liên quan đến các lĩnh vực xuất khẩu, nhập khẩu, hải quan, nhằm phát hiện và kiến nghị xử lý các quy định trái luật, pháp lệnh, trái thẩm quyền, hết hiệu lực, mâu thuẫn, chồng chéo hoặc không còn phù hợp với tình hình mới, gây khó khăn cho hoạt động xuất khẩu, nhập khẩu, cản trở việc thông quan hàng hóa tại các cửa khẩu, các địa điểm

PHAN VĂN KHẢI

QUYẾT ĐỊNH của Thủ tướng Chính phủ số 135/1998/QĐ-TTg ngày 31/07/1998 phê duyệt Chương trình phát triển kinh tế - xã hội các xã đặc biệt khó khăn miền núi và vùng sâu, vùng xa.

THỦ TƯỚNG CHÍNH PHỦ

Căn cứ Luật Tổ chức Chính phủ ngày 30 tháng 9 năm1992;

Căn cứ Nghị quyết số 04/1998/NQ-CP ngày 30/03/1998 của Chính phủ tại phiên họp thường kỳ tháng 3 năm 1998;

Ví dụ về một file PDF có cấu trúc

BẢNG 2: **Một số đặc điểm của các nhóm 'đầu bảng' và 'cuối bảng'**

	TB 53DT	Sán Dìu	Mường	Khmer	Xơ Đăng	Khơ Mú	Mông
Thu nhập TB đầu người ('000 VND)	1.161,4	1.504,3	1.188,9	1.529,4	687,3	511,7	575,2
% ± mức thu nhập trung bình 53DT	0,0	29,5	2,4	31,7	-40,8	-55,9	-50,5
% tỷ lệ nghèo (2015, nghèo thu nhập)	23,1	8,5	18,6	14,8	44,6	59,4	45,7
% cận nghèo (2015, nghèo thu nhập)	13,6	11	19,7	10,3	11,4	13,5	13,4
Quy mô hộ gia đình (người)	4,4	4,1	4,2	4,1	4,4	4,9	5,6
Tuổi thọ bình quân (năm)	72,1	73,22	72,37	72,86	70,5	69,24	68,97
% hộ có nhà kiên cố/bán kiên cố	70,2	72,6	65,4	54,2	80,9	60,1	81,4
% hộ có nhà tạm	15,3	2,8	10,3	38,3	16,3	37,3	14,1
Sử dụng điện lưới quốc gia (%)	93,9	99,9	98,3	98	88,9	58,8	69,9
Sử dụng nước sạch sinh hoạt (%)	73,3	90,7	72,5	93,9	51,1	36,3	53,1
Có nhà xí hợp vệ sinh (%)	27,9	29,6	27,9	36,7	10,3	4,4	7,0
Có xe máy (%)	80,7	89,6	82,3	68,6	59,8	50,9	75,3
Có tủ lạnh (%)	32,2	70,6	44,7	15,5	4,1	4,1	3,6
Có tivi (%)	84,9	95,7	92	87,2	64,5	53,6	50,1

Nguồn: các tác giả tính toán từ Điều tra 53 DTTS.

Trích từ Báo cáo: Các yếu tố ảnh hưởng đến sự phát triển Kinh tế – Xã hội của Dân tộc thiểu số tại Việt Nam - Trang 37, Bảng 2. Nguồn: Worldbank

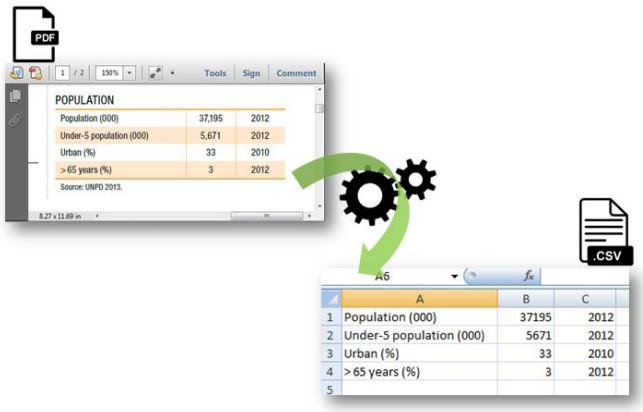
Ví dụ về file PDF có thể thực hiện chức năng tìm kiếm nhưng không có cấu trúc

*Trích từ bài báo khoa học "**Forest resources and forestry in Vietnam**", Journal of Vietnamese Environment,*

https://doi.org/10.13141/jve.vol6.no2

and land use rights for those using land in stable manner or ownership rights over planted production forests. Forestland use rights are 50 years and can be possible extension. Until January 2006, State owned forest and forestland amounted to 10,940,379 ha (76% of total forest and forest land area), households, individuals and the private sector owned forest area calculated to 4,787,762 ha (24% of the total land area) (FAO, 2009). The remaining area belonged to community, cooperative and joint venture enterprises. There were 1,180,465 forestland owners, including 1,173,829 households and individuals, 1,245 commune people's committees, 1,365 economic organizations, 3,105 other entities in addition to a number of enterprises involved in joint ventures with foreign partners and enterprises with foreign investment (Jong et al., 2006; FAO, 2009; MARD, 2007). Changes in forest ownership and utilization have not only been reflected in changes in the structure of forest and land utilization but have also resulted in the establishment of a nationwide

Từ định dạng PDF thành định dạng dữ liệu máy tính có thể đọc được



Dữ liệu ở định dạng PDF

- Các file PDFs có thể chứa các bảng biểu có cấu trúc do máy tính tạo ra nhưng PDF **không phải định dạng dữ liệu** *(data format).*
- Bảng biểu (trong file PDF) cần phải được chuyển đổi sang định dạng có thể mở được bằng phần mềm bảng tính, ví dụ: MS Excel.
- Hay nói cách khác, các bảng dữ liệu này cần được trích xuất thành định dạng dữ liệu thông qua một số phần mềm.

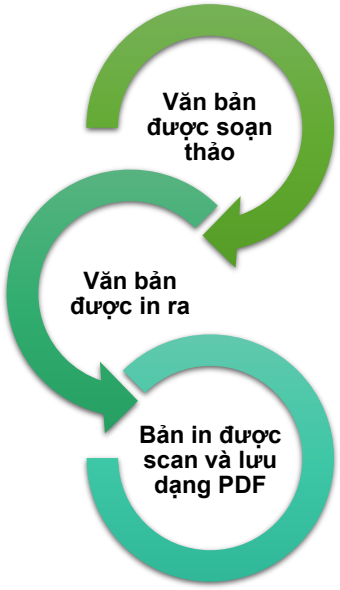
Các công cụ để chuyển đổi file PDF

- Tabula
- Smallpdf
- CometDoc*
- Zamzar*

**have to pay for unlimited conversion per day*

Dữ liệu ở định dạng ảnh scan

- Đây là các file hình ảnh chủ yếu được máy tính đọc dưới dạng một khối hình ảnh thay vì các kí tự riêng biệt.
- Ví dụ: Một số tệp PDF và tất cả các ảnh (GIF, JPEG, PNG, BMP)
- Cần có "Phần mềm nhận dạng ký tự quang học" (*Optical Character Recognition Software,* **OCR**) để nhận dạng kí tự trong file.

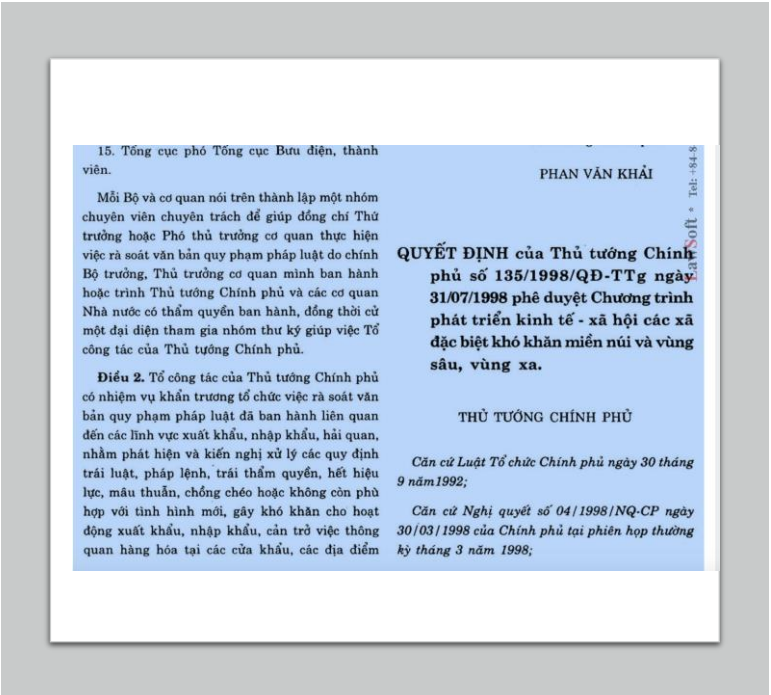


- Dữ liệu được máy tính nhận dạng **theo kí tự**
- Dữ liệu sẽ được máy tính nhận dạng **theo khối hình ảnh**

Ví dụ về một file PDF là ảnh scan

- Quyết định 135/1998/QĐ-TTG phê duyệt Chương trình Phát triển KT-XH các xã đặc biệt khó khăn miền núi và vùng sâu vùng xa

Nguồn: Cổng thông tin Điện tử Chính phủ.



15. Tổng cục phó Tổng cục Bưu điện, thành viên.

Mỗi Bộ và cơ quan nói trên thành lập một nhóm chuyên viên chuyên trách để giúp đồng chí Thứ trưởng hoặc Phó thủ trưởng cơ quan thực hiện việc rà soát văn bản quy phạm pháp luật do chính Bộ trưởng, Thủ trưởng cơ quan mình ban hành hoặc trình Thủ tướng Chính phủ và các cơ quan Nhà nước có thẩm quyền ban hành, đồng thời cử một đại diện tham gia nhóm thư ký giúp việc Tổ công tác của Thủ tướng Chính phủ.

Điều 2. Tổ công tác của Thủ tướng Chính phủ có nhiệm vụ khẩn trương tổ chức việc rà soát văn bản quy phạm pháp luật đã ban hành liên quan đến các lĩnh vực xuất khẩu, nhập khẩu, hải quan, nhằm phát hiện và kiến nghị xử lý các quy định trái luật, pháp lệnh, trái thẩm quyền, hết hiệu lực, mâu thuẫn, chồng chéo hoặc không còn phù hợp với tình hình mới, gây khó khăn cho hoạt động xuất khẩu, nhập khẩu, cản trở việc thông quan hàng hóa tại các cửa khẩu, các địa điểm

PHAN VĂN KHẢI

QUYẾT ĐỊNH của Thủ tướng Chính phủ số 135/1998/QĐ-TTg ngày 31/07/1998 phê duyệt Chương trình phát triển kinh tế - xã hội các xã đặc biệt khó khăn miền núi và vùng sâu, vùng xa.

THỦ TƯỚNG CHÍNH PHỦ

Căn cứ Luật Tổ chức Chính phủ ngày 30 tháng 9 năm1992;

Căn cứ Nghị quyết số 04/1998/NQ-CP ngày 30/03/1998 của Chính phủ tại phiên họp thường kỳ tháng 3 năm 1998;

Các công cụ

- Google Docs OCR
- i2OCR
- Document Cloud

Dữ liệu ở các định dạng có cấu trúc khác

- ***HTML:*** Đây là dữ liệu được hiển thị trên các trang web
- ***JSON:*** Được các nhà lập trình sử dụng rất rộng rãi. Định dạng này có thể xử lý các cấu trúc dữ liệu rất phức tạp
- ***SQL Database:*** Thường được các nhà phân tích dữ liệu và nhà lập trình sử dụng. Không thân thiện với người mới làm quen với dữ liệu nhưng đây là định dạng lưu trữ những bộ dữ liệu cỡ lớn
- ***Geospatial Data:*** GeoJSON, Shapefile, KML, TIFF

Ví dụ về một bảng HTML

Trích từ Trang chuyên đề: ***"Rừng và ngành lâm nghiệp",***

Nguồn: ODV

Mê Kông Campuchia Lào Myanmar Thái Lan **Việt Nam**

Bảng 1: Hiện trạng rừng (Đơn vị: Nghìn Ha)

Xem 25 mục

Năm	Tổng diện tích đất rừng	Rừng tự nhiên	Rừng trồng
2005	12.418,5	9.529,4	2.889,1
2006	12.663,9	10.177,7	2.486,2
2007	12.739,3	10.188,2	2.551,1
2008	13.118,7	10.348,6	2.770,1
2009	13.258,8	10.339,3	2.919,5
2010	13.388,1	10.304,8	3.083,3
2011	13.515,1	10.285,4	3.229,7
2012	13.862	10.423,8	3.438,2
2013	13.954,4	10.398,1	3.556,3
2014	13.796,5	10.100,2	3.693,3
2015	14.061,9	10.175,5	3.886,3
2016	14.377,7	10.242,1	4.135,6

Đang xem 1 đến 12 trong tổng số 12 mục

Ví dụ về cách cấu trúc của tệp JSON

School ID	School Name	Number of Students	Number of Teachers
001	Dagon 1	1200	300
002	Latha 1	1150	200

```
  [
{
  "School ID":"001",
  "School Name": "Dagon 1",
  "Number of Students": 1200,
  "Number of Teachers": 300
},

{
  "School ID":"002",
  "School Name": "Latha 1",
  "Number of Students": 1150,
  "Number of Teachers": 200
}
  ]
```

Bạn có thể thử bằng cách cắt và dán vào đây https://jsoneditoronline.org/

Dữ liệu ở dạng không có cấu trúc

- Một số dữ liệu do máy tính tạo ra nhưng không có cấu trúc để máy có thể nhận dạng được. Ví dụ như các dữ liệu được nhập vào văn bản dưới dạng đoạn văn, hoặc dữ liệu trên website.
- Trong trường hợp này, nhà lập trình cần dạy cho máy tính cách nhận biết pattern (cấu trúc) trong văn bản để có thể trích xuất văn bản thành định dạng dữ liệu.
- Các công cụ: Python hoặc R để chuyển đổi dữ liệu.

Ví dụ về dữ liệu không có cấu trúc

File văn bản word

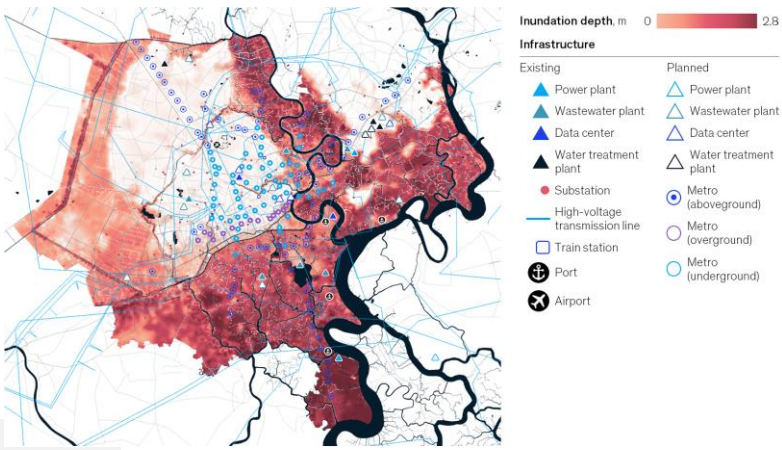
File ảnh(.png, .jpg)

Videos

Audio

Tin trên Social media

Emails



Bản đồ thể hiện trận lụt lịch sử tại TP.HCM trong trường hợp xấu nhất là mực nước biển dâng 180cm - Ảnh: MCKINSEY GLOBAL INSTITUTE

Thank you


This project is funded by the European Union
V 4 M F
Voices for Mekong Forests
OPEN DEVELOPMENT INITIATIVE
AN EAST-WEST MANAGEMENT INSTITUTE PROJECT
liên kết con người và thiên nhiên
PanNature