This project is funded by the European Union

V4MF | Voices for Mekong Forests

Data Literacy Training Phase 1

6-10 July 2020 | Ha Long,Quang Ninh

This project is funded by the European Union

V4MF | Voices for Mekong Forests

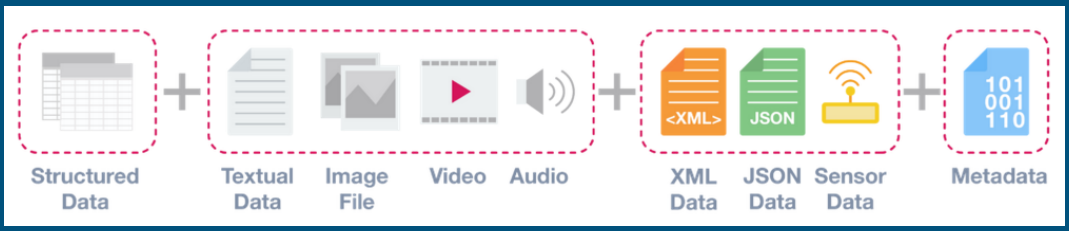
OPEN DEVELOPMENT INITIATIVE
A PROJECT BY EAST-WEST MANAGEMENT INSTITUTE, INC

Tìm kiếm dữ liệu trực tuyến

Sử dụng các công cụ tìm kiếm trực tuyến

Phân loại dữ liệu

- Trong khoa học máy tính, cấu trúc dữ liệu là một cách đặc biệt để tổ chức và lưu trữ dữ liệu trong máy tính sao cho có thể truy cập và sửa đổi một cách hiệu quả. Chính xác hơn, cấu trúc dữ liệu là tập hợp các giá trị dữ liệu, mối quan hệ giữa chúng và các chức năng hoặc hoạt động có thể được áp dụng cho dữ liệu.
- Để phân tích dữ liệu, điều quan trọng là phải hiểu rằng có 3 loại cấu trúc dữ liệu phổ biến: dữ liệu có cấu trúc, dữ liệu phi cấu trúc và dữ liệu bán cấu trúc.



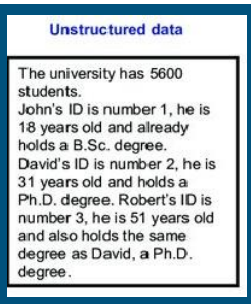
Dữ liệu có cấu trúc

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

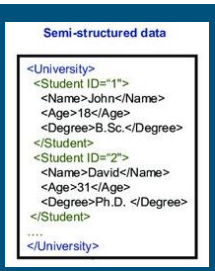
- Dữ liệu có cấu trúc là dữ liệu tuân thủ mô hình dữ liệu được xác định trước và do đó dễ dàng phân tích. Dữ liệu có cấu trúc phù hợp với định dạng bảng có mối quan hệ giữa các hàng và cột khác nhau. Các ví dụ phổ biến của dữ liệu có cấu trúc là các tệp Excel hoặc cơ sở dữ liệu SQL. Mỗi thông tin đều có thể được sắp xếp dưới dạng các hàng và cột.

Dữ liệu phi cấu trúc



- Dữ liệu phi cấu trúc là thông tin không có mô hình dữ liệu được xác định trước hoặc không được tổ chức theo cách được xác định trước. Thông tin phi cấu trúc thường nặng về văn bản, nhưng cũng có thể chứa dữ liệu như ngày, số và sự kiện. Các ví dụ phổ biến về dữ liệu phi cấu trúc bao gồm các tệp âm thanh, video hoặc cơ sở dữ liệu No-SQL.
- Hãy suy nghĩ về hình ảnh, video hoặc tài liệu PDF. Khả năng trích xuất giá trị từ dữ liệu phi cấu trúc là một trong những động lực chính đằng sau sự phát triển nhanh chóng của Dữ liệu lớn.

Dữ liệu bán cấu trúc



- Dữ liệu bán cấu trúc là một dạng dữ liệu có cấu trúc không phù hợp với cấu trúc chính thức của các mô hình dữ liệu được liên kết với cơ sở dữ liệu quan hệ hoặc các dạng bảng dữ liệu khác, nhưng dù sao cũng chứa các thẻ hoặc các dấu khác để phân tách các thành phần ngữ nghĩa và thực thi phân cấp các bản ghi và trường.
- Các ví dụ về dữ liệu bán cấu trúc như:
  - Ngôn ngữ văn bản XML
  - Định dạng file JSON

Metadata: Dữ liệu về dữ liệu

- Một loại cuối cùng của loại dữ liệu là dữ liệu mô tả. Từ quan điểm kỹ thuật, đây không phải là một cấu trúc dữ liệu riêng biệt, nhưng nó là một trong những yếu tố quan trọng nhất để phân tích Dữ liệu lớn và các giải pháp dữ liệu lớn. Metadata là dữ liệu về dữ liệu. Nó cung cấp thông tin bổ sung về một bộ dữ liệu cụ thể.

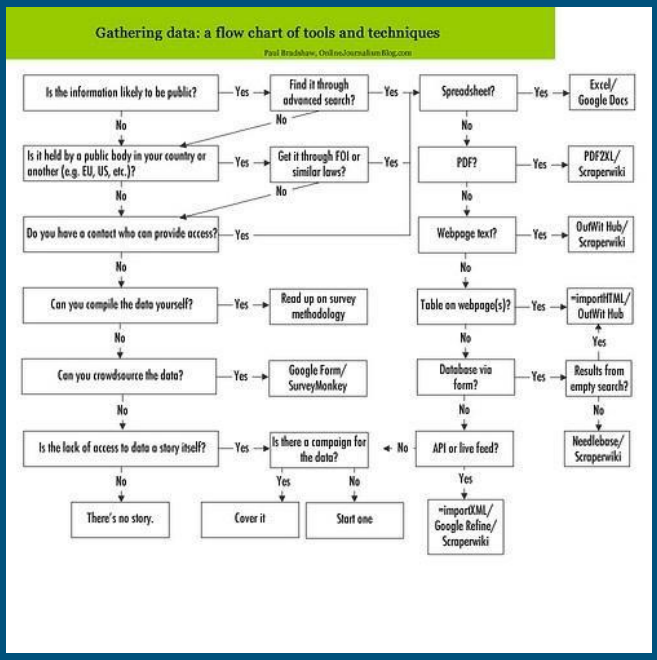
Sensor	Date (mm-dd-yyyy)	Path/row	Resolution/m
MSS	10-28-1973	145/33	57
TM	8-19-1995	135/33	30
ETM+	16-8-1999	135/33	15
TM	8-1-2006	135/33	30
TM	9-13-2010	135/33	30

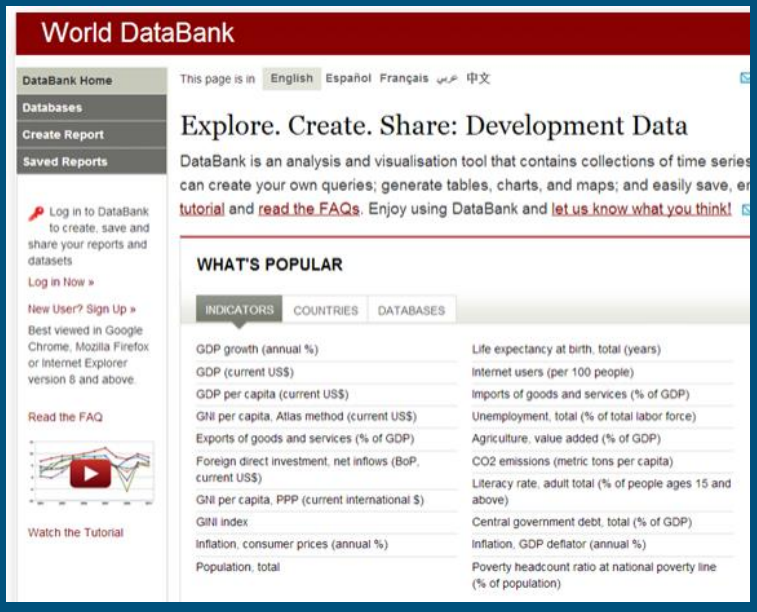
doi:10.1371/journal.pone.0070574.t001

Chìm đắm trong dữ liệu

Trong kỷ nguyên số, có nhiều dữ liệu sẵn có hơn bao giờ hết. Trên thực tế, đôi khi cảm giác như chúng ta đang bị chìm đắm trong bể dữ liệu nhưng lại rất khó để tìm được dữ liệu mà chúng ta thực sự muốn.

Trong bài này, chúng ta sẽ tìm hiểu cách để tìm dữ liệu trực tuyến thông qua các cổng thông tin và các công cụ tìm kiếm.





https://data.worldbank.org/

Định hướng dùng World Databank

Khi nhìn vào trang này, có rất nhiều biệt ngữ về dữ liệu có thể khiến ta nhầm lẫn. Một số các hạng mục đầu tiên như, tăng trưởng GDP như thế nào (%hàng năm); GDP (giá hiện tại US$); GNI đầu người, phương pháp Atlas (giá hiện tại) và Xuất khẩu hàng hóa và dịch vụ (% của GDP) – có nghĩa là gì. Chúng ta không thể mong mọi người dùng dữ liệu kinh tế trong báo cáo đều là các nhà kinh tế học.

Thay vào đó, đây là nơi này sẽ giúp bạn định hướng xung quanh các dữ liệu phức tạp bằng việc không chỉ giới thiệu một số định dạng dữ liệu cơ bản và các câu hỏi chúng ta luôn luôn nên hỏi về bộ dữ liệu, mà còn các tài liệu mà chúng ta cần tìm ra những gì vượt trên kiến thức của chúng ta giúp chúng ta hiểu dữ liệu đang đo lường cái gì và như thế nào.

Lời khuyên để tìm đúng dữ liệu

- Nghiên cứu đề tài của mình để biết ai đang thu thập dữ liệu về chủ đề đó
- Thực hiện thật nhiều tìm kiếm trên Google!
- Đọc các chú thích, tài liệu tham khảo và định nghĩa
- Giả định rằng dữ liệu có sẵn ở đâu đó

Danh sách một số địa chỉ có thể tìm kiếm dữ liệu

- https://toolbox.google.com/datasetsearch
- https://www.kaggle.com/datasets
- https://github.com/search?q=datasets
- https://github.com/awesomedata/awesome-public-datasets
- https://www.data.gov/
- https://www.usa.gov/statistics
- https://www.federalreserve.gov/data.htm
- https://www.bls.gov/
- https://data.ca.gov/
- https://data.fivethirtyeight.com/
- https://www.opendatacube.org/ceos
- https://data.world/search

Danh sách một số địa chỉ có thể tìm kiếm dữ liệu

- https://data.oecd.org/
- https://openmaptiles.com/
- https://www.openstreetmap.org
- Dữ liệu về nông nghiệp
- Dữ liệu về mô hình số độ cao
- Dữ liệu ảnh viễn thám
- Dữ liệu ảnh Jaxa
- Dữ liệu ảnh Lidar
- Dữ liệu DEM của MOLA
- Các view mới cập nhật của NASA, ESA, JAXA về COVID-19
- .....

Danh sách một số webiste có dữ liệu tại Việt Nam

- https://opendevelopmentvietnam.net/
- https://opendata.hochiminhcity.gov.vn/search/type/dataset
- https://congdulieu.vn/
- https://portal.mrcmekong.org/home
- https://www.gso.gov.vn/Default.aspx?tabid=217
- https://dulieu.itrithuc.vn/dataset
- http://dvmtr.siteam.vn/TraCuu
- http://rungvenbien.ifee.edu.vn/ThongKeBaoCao/Index
- http://dktb.dichvucong.vinamarine.gov.vn/WebDKTB/TraCuuDuLieuTau.aspx

Tìm kiếm nâng cao trên Google

10 mẹo để tìm kiếm Internet thông minh hơn, hiệu quả hơn

1. Sử dụng thuật ngữ độc đáo, cụ thể
2. Sử dụng toán tử (-) để thu hẹp tìm kiếm
3. Sử dụng dấu ngoặc kép cho các cụm từ chính xác
4. Không sử dụng các từ phổ biến và dấu câu (từ loại chung)
5. Không phân biệt viết hoa hay viết thường
6. Bỏ các từ phụ (số nhiều, các từ ngữ đi kèm)
7. Tối đa hóa sự gợi ý
    a. https://www.google.com/
    b. https://www.google.com/webhp?complete=0
8. Tùy chỉnh tìm kiếm của bạn (Sử dụng các toán tử và tìm kiếm nâng cao)
9. Sử dụng lịch sử trình duyệt

10 mẹo để tìm kiếm Internet thông minh hơn, hiệu quả hơn

10. Đặt giới hạn thời gian - sau đó thay đổi chiến thuật

Đôi khi, bạn không bao giờ có thể tìm thấy những gì bạn đang tìm kiếm. Bắt đầu một đồng hồ nội bộ, và khi một khoảng thời gian nhất định đã trôi qua mà không có kết quả, hãy ngừng đập đầu vào tường. Đã đến lúc thử một thứ khác:

a. Sử dụng một công cụ tìm kiếm khác, như Yahoo!, Bing, Startpage hoặc Lycos.
b. Hỏi một người ngang hàng.
c. Gọi hỗ trợ.
d. Đặt một câu hỏi trong diễn đàn thích hợp.
e. Sử dụng các chuyên gia hoặc tìm kiếm những người có thể tìm thấy câu trả lời cho bạn.

Tìm kiếm Internet với một số toán tử phổ biến

1. OR (|) , AND , - , ( ), $,
2. define: , cache:, filetype: (ext:)
3. site: , related: , intitle: , allintitle: , inurl: , alllinurl: ,
4. intext:, allintext: ,
5. AROUND(X), weather: ,
6. stock:, map:, movie:, source:,
7. blogurl:, inanchor:, allinanchor:, location:

Advanced google search