Muriel Foulonneau

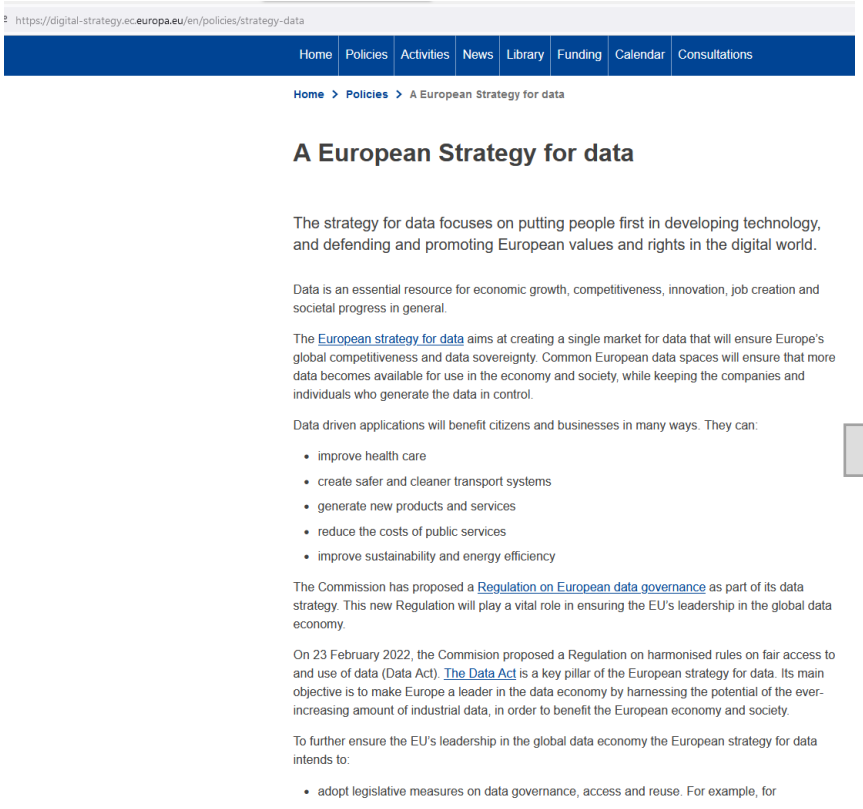
Data quality as a risk

Stories of data reuse

ENDORSE

THE EUROPEAN DATA CONFERENCE ON REFERENCE DATA AND SEMANTICS

The promise of the data economy



https://digital-strategy.ec.europa.eu/en/policies/strategy-data

Home | Policies | Activities | News | Library | Funding | Calendar | Consultations

Home > Policies > A European Strategy for data

A European Strategy for data

The strategy for data focuses on putting people first in developing technology, and defending and promoting European values and rights in the digital world.

Data is an essential resource for economic growth, competitiveness, innovation, job creation and societal progress in general.

The European strategy for data aims at creating a single market for data that will ensure Europe's global competitiveness and data sovereignty. Common European data spaces will ensure that more data becomes available for use in the economy and society, while keeping the companies and individuals who generate the data in control.

Data driven applications will benefit citizens and businesses in many ways. They can:

- improve health care
- create safer and cleaner transport systems
- generate new products and services
- reduce the costs of public services
- improve sustainability and energy efficiency

The Commission has proposed a Regulation on European data governance as part of its data strategy. This new Regulation will play a vital role in ensuring the EU's leadership in the global data economy.

On 23 February 2022, the Commision proposed a Regulation on harmonised rules on fair access to and use of data (Data Act). The Data Act is a key pillar of the European strategy for data. Its main objective is to make Europe a leader in the data economy by harnessing the potential of the ever-increasing amount of industrial data, in order to benefit the European economy and society.

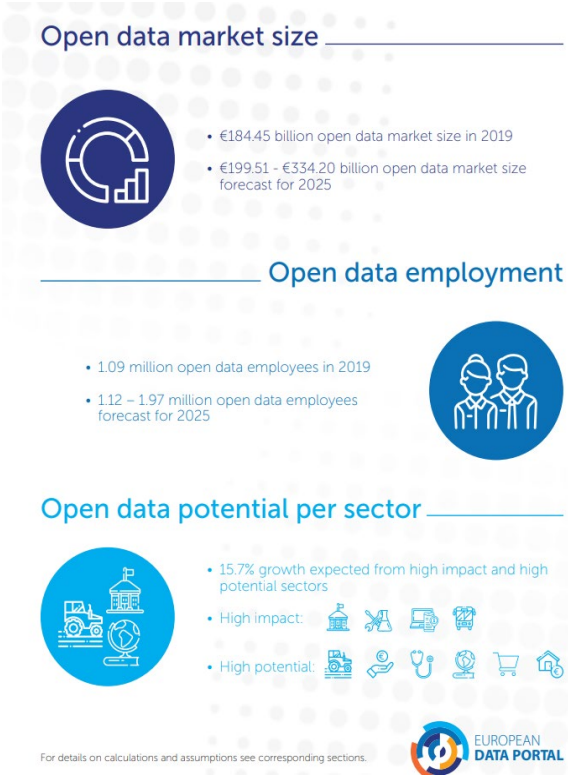
To further ensure the EU's leadership in the global data economy the European strategy for data intends to:

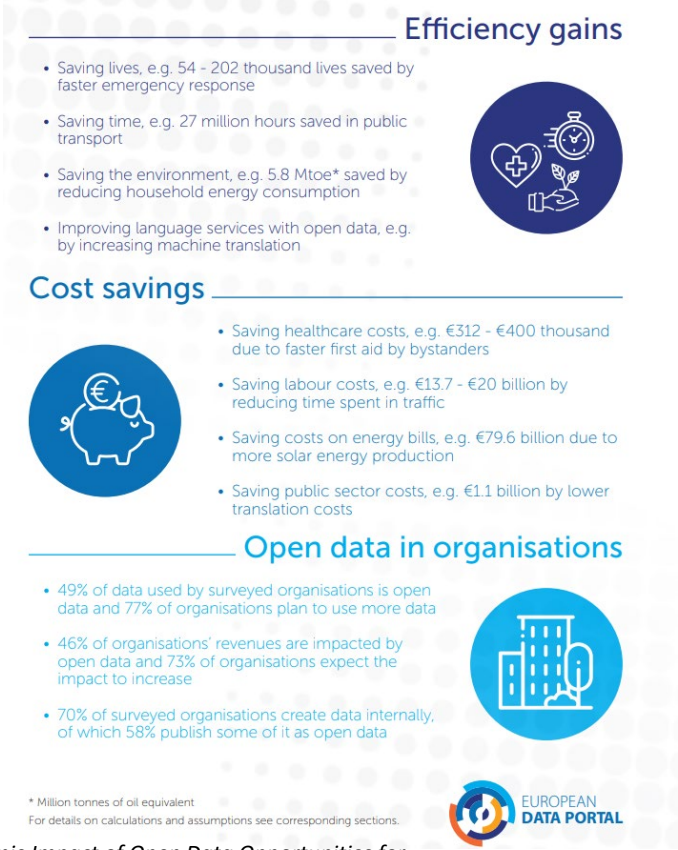
- adopt legislative measures on data governance, access and reuse. For example, for

Data creates value

Sharing data creates more value

Reusing data to support innovation



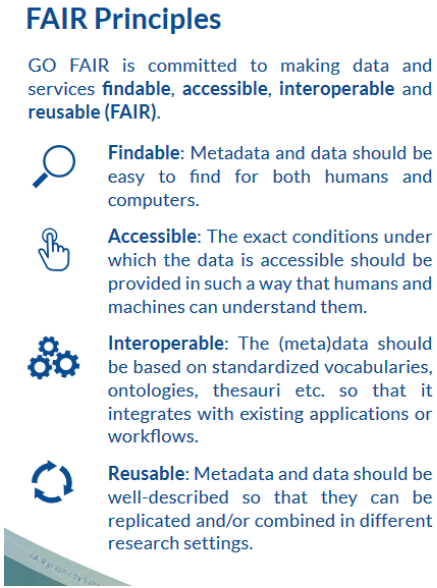


ENDORSE

Esther Huyer & Laura van Knippenberg. The Economic Impact of Open Data Opportunities for value creation in Europe. Etude pour le European Data Portal. 2020

What is reusable data?

- Data reusers raise data quality as a major obstacle
  - Ex. Etalab survey 2020:
    - Issues on freshness
    - Insufficient or inaccurate documentation
    - Issues on dataset uniqueness



https://www.go-fair.org/wp-content/uploads/2020/02/GO_FAIR_Flyer_Update19022020-1.pdf

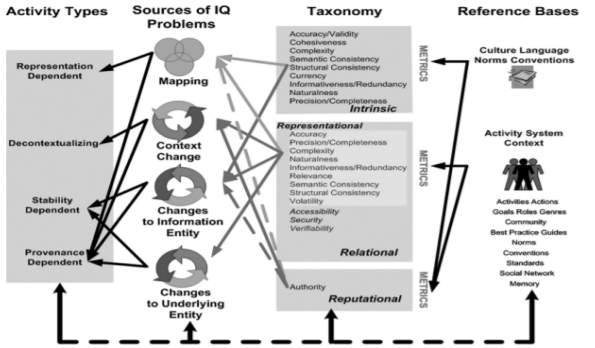
How can we measure data quality?

- Multiple frameworks and dimensions

European data portal/JoinUp

- **Accuracy**: is the data correctly representing the real-world entity or event?
- **Consistency**: Is the data not containing contradictions?
- **Availability**: Can the data be accessed now and over time?
- **Completeness**: Does the data include all data items representing the entity or event?
- **Conformance**: Is the data following accepted standards?
- **Credibility**: Is the data based on trustworthy sources?
- **Processability**: Is the data machine-readable?
- **Relevance** : Does the data include an appropriate amount of data?
- **Timeliness**: Is the data representing the actual situation and is it published soon enough?

Metadata, data and information quality



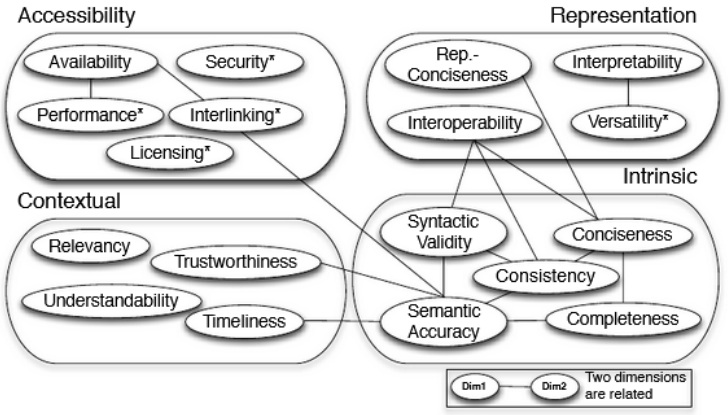
Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. Journal of the American society for information science and technology, 58(12), 1720-1733

IMF

TABLE 1 The Six Dimensions of Data Quality

Relevance	The *relevance* of statistical information reflects the degree to which it meets the real needs of clients. It is concerned with whether the available information sheds light on the issues of most importance to users. Assessing relevance is a subjective matter dependent upon the varying needs of users. The NSO's challenge is to weigh and balance the conflicting needs of different users to produce a program that goes as far as possible in satisfying the most important needs and users within given resource constraints.
Accuracy	The *accuracy* of statistical information is the degree to information correctly describes the phenomena it was d measure. It is usually characterized in terms of error in estimates and is traditionally decomposed into bias (sys variance (random error) components. It may also be des of the major sources of error that potentially cause inacc coverage, sampling, nonresponse, response).
Timeliness	The *timeliness* of statistical information refers to the de reference point (or the end of the reference period) to wh information pertains, and the date on which the informa available. It is typically involved in a trade-off against *a timeliness* of information will influence its *relevance*.
Accessibility	The *accessibility* of statistical information refers to the can be obtained from the NSO. This includes the ease w existence of information can be ascertained, as well as t the form or medium through which the information can cost of the information may also be an aspect of *accessi* users.
Interpretability	The *interpretability* of statistical information reflects the supplementary information and metadata necessary to in utilize it appropriately. This information normally cover concepts, variables and classifications used, the method collection, and indications of the accuracy of the statisti
Coherence	The *coherence* of statistical information reflects the deg can be successfully brought together with other statistic within a broad analytic framework and over time. The u concepts, classifications and target populations promote does the use of common methodology across surveys. C not necessarily imply full numerical consistency.

Linked data



Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. Semantic Web, 7(1), 63-93.

A study in 2020 by DAMA NL found 127 quality dimensions

Nr	Dimensions of data quality	Classification	Source
1.	Ease of operation		
2.	Reproducibility		
3.	Granularity		-
4.	Retention period		-
5.	Accessibility		CDDQ 2019
6.	Accuracy		CDDQ 2019
7.	Completeness		CDDQ 2019
8.	Consistency		CDDQ 2019
9.	Currency		CDDQ 2019
10.	Integrity		CDDQ 2019
11.	Lineage		CDDQ 2019
12.	Precision		CDDQ 2019
13.	Representation		CDDQ 2019
14.	Timeliness		CDDQ 2019
15.	Validity		CDDQ 2019
16.	Coverage		Daas 2010
17.	Likability		Daas 2010
18.	Accuracy		DAMA 2017
19.	Completeness		DAMA 2017
20.	Consistency		DAMA 2017
21.	Currency (of data)	Timeliness	DAMA 2017
22.	Integrity or Coherence		DAMA 2017
23.	Latency	Timeliness	DAMA 2017
24.	Reasonability		DAMA 2017
25.	Timeliness		DAMA 2017
26.	Uniqueness		DAMA 2017
27.	Validity		DAMA 2017
28.	Volatility	Timeliness	DAMA 2017
29.	Accuracy	Core dimension	DAMA-UK 2013
30.	Completeness	Core dimension	DAMA-UK 2013
31.	Confidence		DAMA-UK 2013
32.	Consistency	Core dimension	DAMA-UK 2013
33.	Flexibility		DAMA-UK 2013
34.	Timeliness	Core dimension	DAMA-UK 2013
35.	Uniqueness	Core dimension	DAMA-UK 2013
36.	Usability		DAMA-UK 2013
37.	Validity	Core dimension	DAMA-UK 2013
38.	Value		DAMA-UK 2013
39.	Accessibility	Pragmatic	English 1999
40.	Accuracy		English 1999
41.	Accuracy to a surrogate source	Inherent	English 1999
42.	Completeness (of values)	Inherent	English 1999
43.	Concurrency (of redundant or distributed data)	Inherent	English 1999
44.	Contextual clarity	Pragmatic	English 1999
45.	Database integrity		English 1999
46.	Definition conformance (see metadata conformance)	Inherent	English 1999
47.	Entity integrity		English 1999
48.	Equivalence	Inherent	English 1999
49.	Fact completeness	Pragmatic	English 1999
50.	Flexibility		English 1999
51.	Non-duplicates	Inherent	English 1999
52.	Precision	Inherent	English 1999
53.	Stability		English 1999
54.	Timeliness	Pragmatic	English 1999
55.	Usability	Pragmatic	English 1999
56.	Validity	Inherent	English 1999
57.	Accessibility		Eurostat 2015
58.	Accuracy		Eurostat 2015
59.	Clarity		Eurostat 2015
60.	Coherence		Eurostat 2015
61.	Comparability		Eurostat 2015
62.	Confidentiality		Eurostat 2015
63.	Consistency		Eurostat 2015
64.	Punctuality		Eurostat 2015
65.	Relevance		Eurostat 2015
66.	Reliability		Eurostat 2015
67.	Timeliness		Eurostat 2015
68.	Accessibility	Inherent/System dependant	ISO 25012
69.	Accuracy	Inherent	ISO 25012
70.	Availability	System dependant	ISO 25012
71.	Completeness	Inherent	ISO 25012
72.	Compliance	Inherent/System dependant	ISO 25012
73.	Confidentiality	Inherent/System dependant	ISO 25012
74.	Consistency	Inherent	ISO 25012
75.	Credibility	Inherent	ISO 25012
76.	Currentness	Inherent	ISO 25012
77.	Efficiency	Inherent/System dependant	ISO 25012
78.	Portability	System dependant	ISO 25012
79.	Precision	Inherent/System dependant	ISO 25012
80.	Recoverability	System dependant	ISO 25012
81.	Traceability	Inherent/System dependant	ISO 25012
82.	Understandability	Inherent/System dependant	ISO 25012
83.	Ability to represent null values	Representation	Redman 1996
84.	Accordance with format (of the physical instances)	Representation	Redman 1996
85.	Accuracy	Data values	Redman 1996
86.	Appropriateness	Representation	Redman 1996
87.	Clarity	Content	Redman 1996
88.	Completeness	Data values	Redman 1996
89.	Consistency	Data values	Redman 1996
90.	Currency	Data values	Redman 1996
91.	Efficient use (of storage)	Representation	Redman 1996
92.	Flexibility	Reaction to change	Redman 1996
93.	Format flexibility	Representation	Redman 1996
94.	Format precision	Representation	Redman 1996
95.	Granularity (of attributes)	Level of detail	Redman 1996
96.	Homogeneity	Composition	Redman 1996
97.	Identify-ability	Composition	Redman 1996
98.	Interpretability	Representation	Redman 1996
99.	Level of detail	Level of detail	Redman 1996
100.	Naturalness	Composition	Redman 1996
101.	Obtainability	Content	Redman 1996
102.	Portability	Representation	Redman 1996
103.	Precision (of attribute domains)	Level of detail	Redman 1996
104.	Redundancy (minimum necessary)	Composition	Redman 1996
105.	Relevance	Content	Redman 1996
106.	Robustness	Reaction to change	Redman 1996
107.	Scope	Scope	Redman 1996
108.	Semantic consistency (of the components of the model)	View consistency	Redman 1996
109.	Structural consistency (of attributes across entity types)	View consistency	Redman 1996
110.	Access security	Accessibility	Wang 1996
111.	Accessibility	Accessibility	Wang 1996
112.	Accuracy	Intrinsic	Wang 1996
113.	Appropriateness (of amount of data)	Contextual	Wang 1996
114.	Believability	Intrinsic	Wang 1996
115.	Completeness	Contextual	Wang 1996
116.	Conciseness (of representation)	Representational	Wang 1996
117.	Cost-effectiveness		Wang 1996
118.	Ease of understanding	Representational	Wang 1996
119.	Interpretability	Representational	Wang 1996
120.	Objectivity	Intrinsic	Wang 1996
121.	Relevancy	Contextual	Wang 1996
122.	Representational consistency	Representational	Wang 1996
123.	Reputation	Intrinsic	Wang 1996
124.	Timeliness	Contextual	Wang 1996
125.	Traceability		Wang 1996
126.	Value-added	Contextual	Wang 1996
127.	Variety		Wang 1996

ENDORSE

https://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf

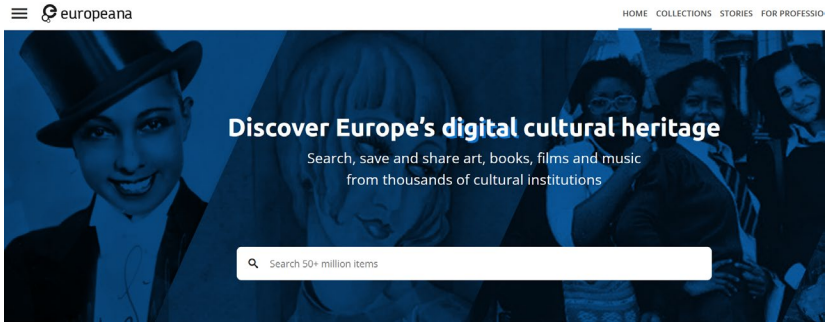
A few examples of data reuse

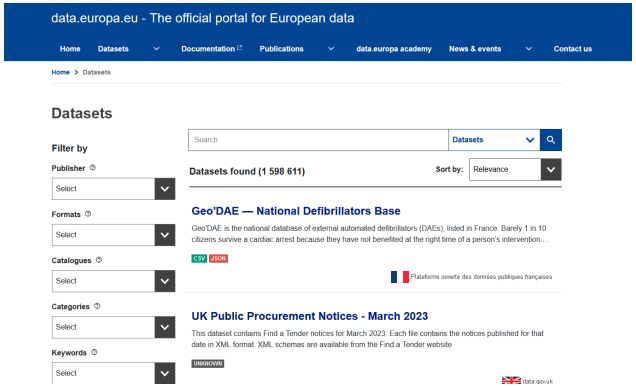
and the data quality issues they raised

A distributed digital library

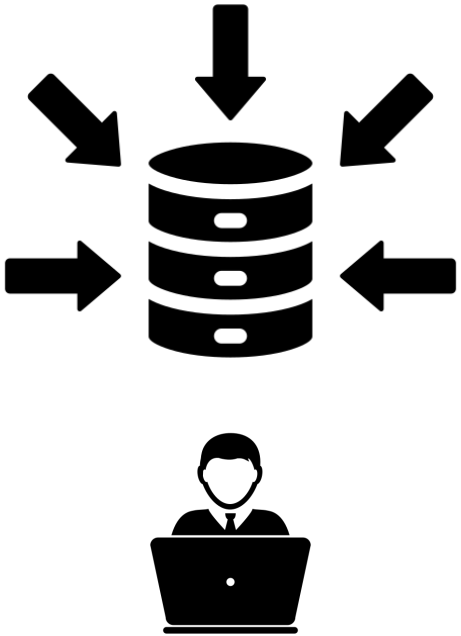
ENDORSE

Metadata aggregation





Agregate and share: Low cost, low touch



Decontextualized data

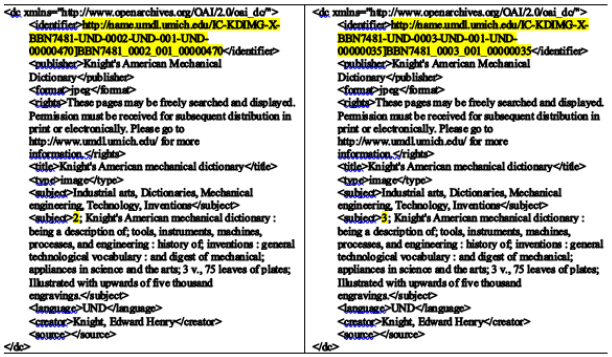
As Colonel of the Rough Riders, 1898.

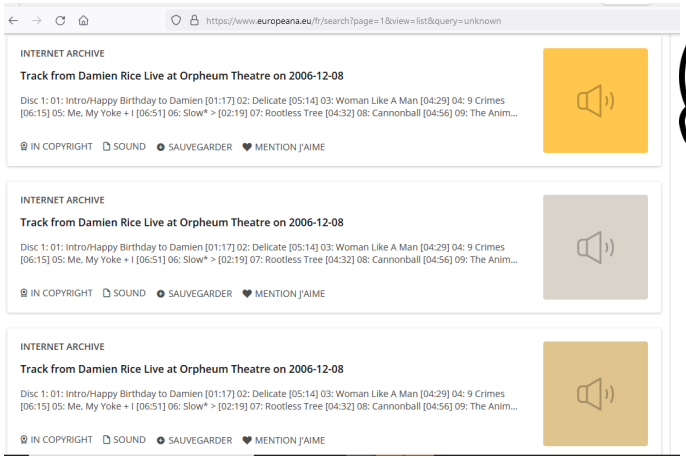
https://hollis.harvard.edu/primo-explore/fulldisplay?docid=HVD_VIAolvgroup12088&context=L&vid=HVD2&search_scope=everything&tab=everything&lang=en_US

- By using the collection description we could create a full match on 17% of multi-term searches

Foulonneau, M., Cole, T. W., Habing, T. G., & Shreeves, S. L. (2005). Using collection descriptions to enhance an aggregation of harvested item-level metadata. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 2005. JCDL'05, (pp. 32 –41).

Wendler, Robin. The Eye of the Beholder: Challenges of Image Description and Access at Harvard. In Hillmann, Diane I. and Westbrooks, Elaine L., eds., Metadata in Practice. American Library Association, Chicago, IL, 2004, 51-69.





Pseudo-duplicates

```
<dc xmlns="http://www.openarchives.org/OAI/2.0/oai_dc/">
    <identifier>http://name.umdl.umich.edu/IC-KDIMG-X-
    BBN7481-UND-0002-UND-001-UND-
    00000470]BBN7481_0002_001_00000470</identifier>
    <publisher>Knight's American Mechanical
    Dictionary</publisher>
    <format>jpeg</format>
    <rights>These pages may be freely searched and displayed.
    Permission must be received for subsequent distribution in
    print or electronically. Please go to
    http://www.umdl.umich.edu/ for more
    information.</rights>
    <title>Knight's American mechanical dictionary</title>
    <type>image</type>
    <subject>Industrial arts, Dictionaries, Mechanical
    engineering, Technology, Inventions</subject>
    <subject>2; Knight's American mechanical dictionary :
    being a description of; tools, instruments, machines,
    processes, and engineering : history of; inventions : general
    technological vocabulary : and digest of mechanical;
    appliances in science and the arts; 3 v., 75 leaves of plates;
    Illustrated with upwards of five thousand
    engravings.</subject>
    <language>UND</language>
    <creator>Knight, Edward Henry</creator>
    <source></source>
</dc>
```

```
<dc xmlns="http://www.openarchives.org/OAI/2.0/oai_dc/">
    <identifier>http://name.umdl.umich.edu/IC-KDIMG-X-
    BBN7481-UND-0003-UND-001-UND-
    00000035]BBN7481_0003_001_00000035</identifier>
    <publisher>Knight's American Mechanical
    Dictionary</publisher>
    <format>jpeg</format>
    <rights>These pages may be freely searched and displayed.
    Permission must be received for subsequent distribution in
    print or electronically. Please go to
    http://www.umdl.umich.edu/ for more
    information.</rights>
    <title>Knight's American mechanical dictionary</title>
    <type>image</type>
    <subject>Industrial arts, Dictionaries, Mechanical
    engineering, Technology, Inventions</subject>
    <subject>3; Knight's American mechanical dictionary :
    being a description of; tools, instruments, machines,
    processes, and engineering : history of; inventions : general
    technological vocabulary : and digest of mechanical;
    appliances in science and the arts; 3 v., 75 leaves of plates;
    Illustrated with upwards of five thousand
    engravings.</subject>
    <language>UND</language>
    <creator>Knight, Edward Henry</creator>
    <source></source>
</dc>
```

https://www.europeana.eu/fr/search?page=1&view=list&query=unknown

INTERNET ARCHIVE

Track from Damien Rice Live at Orpheum Theatre on 2006-12-08

Disc 1: 01: Intro/Happy Birthday to Damien [01:17] 02: Delicate [05:14] 03: Woman Like A Man [04:29] 04: 9 Crimes [06:15] 05: Me, My Yoke + I [06:51] 06: Slow* > [02:19] 07: Rootless Tree [04:32] 08: Cannonball [04:56] 09: The Anim...

INTERNET ARCHIVE

Track from Damien Rice Live at Orpheum Theatre on 2006-12-08

Disc 1: 01: Intro/Happy Birthday to Damien [01:17] 02: Delicate [05:14] 03: Woman Like A Man [04:29] 04: 9 Crimes [06:15] 05: Me, My Yoke + I [06:51] 06: Slow* > [02:19] 07: Rootless Tree [04:32] 08: Cannonball [04:56] 09: The Anim...

INTERNET ARCHIVE

Track from Damien Rice Live at Orpheum Theatre on 2006-12-08

Disc 1: 01: Intro/Happy Birthday to Damien [01:17] 02: Delicate [05:14] 03: Woman Like A Man [04:29] 04: 9 Crimes [06:15] 05: Me, My Yoke + I [06:51] 06: Slow* > [02:19] 07: Rootless Tree [04:32] 08: Cannonball [04:56] 09: The Anim...

Paris autrement - Balades dans les arrondissements périphériques - POI

Balades insolites au cœur des 12, 13 et 14ème arrondissements - Jeu de données présentant des parcours et des points d'intérêts sur les trésors cachés de ces arrondissements.

Modifié	30 octobre 2019 08:50
Producteur	Direction de l'Attractivité et de l'Emploi - Ville de Paris
Licence	Open Database License (ODbL)
Mots clés	tourisme, balade
Données	157 éléments

Paris autrement - Balades dans les arrondissements périphériques - Parcours

Balades insolites au cœur des 12, 13 et 14ème arrondissements - Jeu de données présentant des parcours et des points d'intérêts sur les trésors cachés de ces arrondissements.

Modifié	30 octobre 2019 08:48
Producteur	Direction de l'Attractivité et de l'Emploi - Ville de Paris
Licence	Open Database License (ODbL)
Mots clés	tourisme, balade
Données	12 éléments

https://www.data.gouv.fr

- For 2.8% of collections, every time a query matched an item of the collection, it matched all of them

Foulonneau, M. (2007). Information redundancy across metadata collections. Information processing & management, 43(3), 740-751.

Standardization: Data does not always conform to expectations

DC Date

September 29–October 28, 51 AD; 1970
second half of IXth century AD; 1978
Rebuilt 1984
Possibly Vth/VIth century AD; 1935
Planted 1985
n/a
n.d.
Mid IInd century AD; 1973
Jul-51
circa 900 AD
ca. 701 BC
Begun 14th century
184-?
1839
18–?
August 23, 2000
between 1827 and 183
VIIIth/IXth century AD ? (TC);1965
Vth-VIth century AD (McNamee); IVth century AD (Cribiore); 1982
XVIII Dynasty
Winter 2003
era of redevelopment
various
2002-00
1980, refurbished 1997
China: Neolithic Period (5000 BCE-ca 1600 BCE)?
19691968
21. Nouemb. Anno. 1564.
And finisshed on the euen of thanunciacion of our said bilissid Lady falling on the wednesday the xxiiij daye of Marche. in the xix yeer of Kyng Edwarde the fourthe
[1479]]
19193
xxxx Oct xx
Various
1938-05-38
1963 to 1953
[not after 1579]
163[5?]

Dimension (width x height)	Descriptive name
48x48	very small
64x64	Small
96x96	Medium
128x128	Large
144x144	extra large
160x160	super large
192x192	ridiculous large

Jens Finke's reference sizes for thumbnails

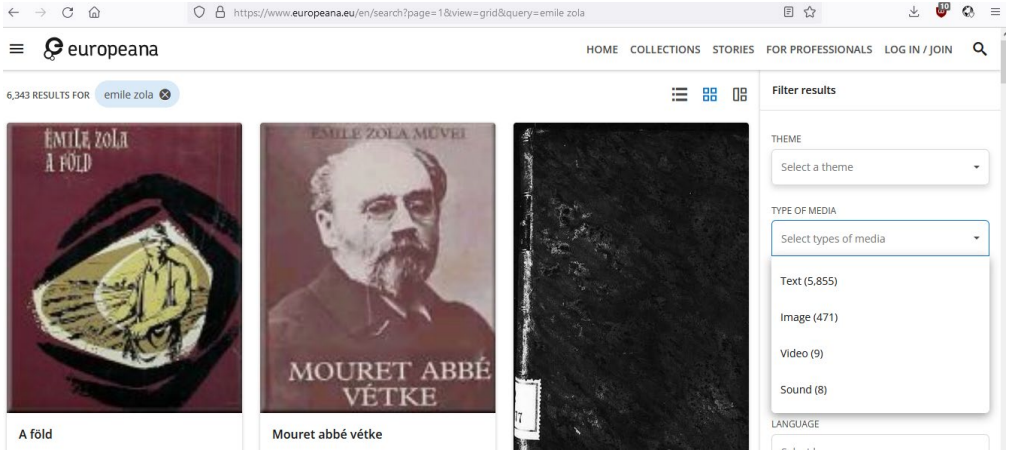
006

ENDORSE

Data completeness and usability for specific user tasks

Dublin Core element	% of repositories using element at least once	No. of records containing element	Total times element used	% of total records containing element	Average times used per record	Average element length (in characters)	Mode	Mode Frequency in %
Title	100.0	124,304	133,108	80.3	1.1	39.9	1	75.8
Creator	87.5	78,402	84,829	50.7	1.1	21.5	0	49.3
Subject	93.8	112,875	304,661	72.9	2.7	110.4	2	37.1
Description	81.3	73,298	153,088	47.4	2.1	104.1	0	52.6
Publisher	75.0	94,791	114,305	61.2	1.2	38.5	1	50.9
Contributor	62.5	10,158	16,813	6.6	1.7	47.0	0	93.4
Date	81.3	66,514	77,175	43.0	1.2	10.9	0	57.0
Type	81.3	118,419	124,853	76.5	1.1	6.6	1	72.5
Format	56.3	107,381	111,647	69.4	1.0	8.3	1	66.6
Identifier	100.0	154,113	205,719	99.6	1.3	84.4	1	71.5
Source	50.0	23,012	29,537	14.9	1.3	68.3	0	85.1
Language	75.0	85,201	85,397	55.0	1.0	3.3	1	54.9
Relation	43.8	48,356	80,629	31.2	1.7	95.6	0	68.8
Coverage	37.5	9,136	12,103	5.9	1.3	21.0	0	94.1
Rights	62.5	63,435	68,228	41.0	1.1	151.7	0	59.0

Stvilia, B., Gasser, L., & Twidale, M. B. (2007). Metadata quality problems in federated collections. In Challenges of Managing Information Quality in Service Organizations (pp. 154-186). IGI Global.



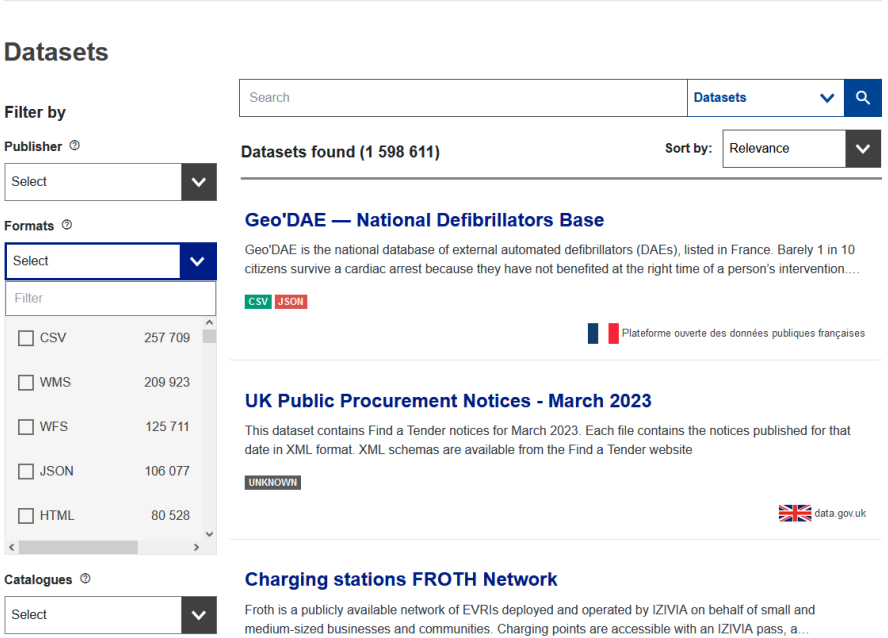
ENDORSE

Assess and show impact based on user tasks

User tasks

- Find
- Identify
- Select
- Obtain
- Explore

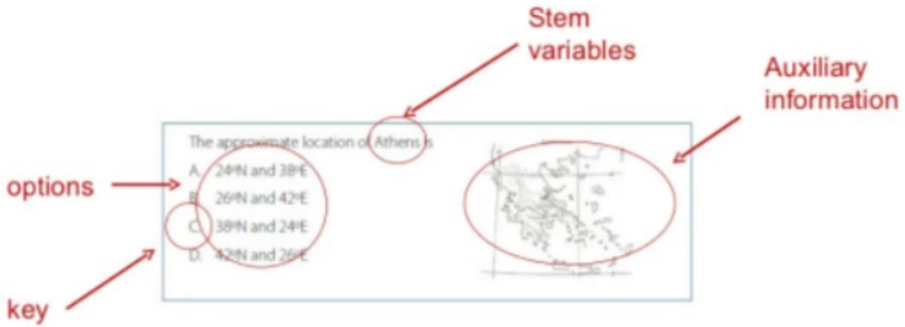
(FRSAD and FRBR)



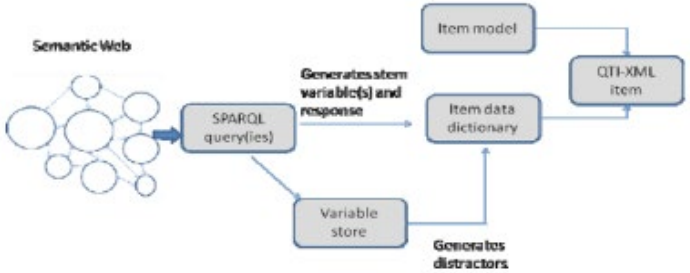
ENDORSE

Educational test items generation

Assessment item generation



Foulonneau, M., Ras, E. (2013). Assessment Item Generation, the way forward. International Computer Assisted Assessment (CAA) Conference, Southampton, UK
Adapted from Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). Automatic item generation: Theory and practice. Routledge. New York



Foulonneau, M. (2012). Generating educational assessment items from linked open data: The case of DBpedia. In Extended Semantic Web Conference (pp. 16-27). Springer, Berlin, Heidelberg.

Risks on datasets interdependencies

```
<?xml version="1.0" encoding="UTF-8" standalone="no"
- <assessmentItem xmlns="http://www.imsglobal.org/
  instance" adaptive="false" identifier="choice" timeDe
  xsi:schemaLocation="http://www.imsglobal.org/x
  - <responseDeclaration baseType="identifier" cardinalit
    - <correctResponse>
        <value>Option2</value>
      </correctResponse>
    </responseDeclaration>
  - <outcomeDeclaration baseType="integer" cardinality=
    - <defaultValue>
        <value>0</value>
      </defaultValue>
    </outcomeDeclaration>
  - <itemBody>
    - <p>
        <img alt="flag" height="20" src="http://upload.wikimedia.org/wikipedia/commons/d/da/Flag_of_Luxembourg.svg" />
      </p>
    - <choiceInteraction maxChoices="1" responseIdentifier="RESPONSE" shuffle="false">
        <prompt>Which country is represented by this flag ?</prompt>
        <simpleChoice identifier="Option0">Bulgaria</simpleChoice>
        <simpleChoice identifier="Option1">Azerbaijan</simpleChoice>
        <simpleChoice identifier="Option2">Luxembourg</simpleChoice>
      </choiceInteraction>
    </itemBody>
  <responseProcessing template="http://www.imsglobal.org/question/qti_v2p0/rptemplates/match_correct" />
</assessmentItem>
```

About: Luxembourg

An Entity of Type: musicien, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Luxemburg est une ville du comté de Kewaunee dans le Wisconsin. Sa population était de 2 515 habitants en 2010.

Property	Value
dbo:PopulatedPlace/area	• 2584.808134115328 • 2586.4
dbo:PopulatedPlace/populationDensity	• 242.0 • 232.7423811693864

- 6 out of 30 missing links

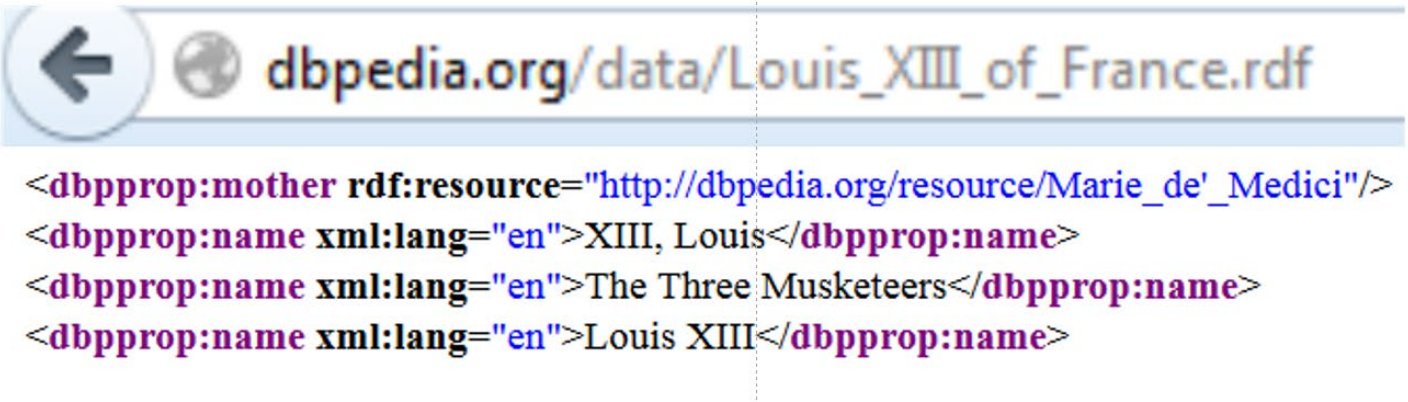
Foulonneau, M. (2012). Generating educational assessment items from linked open data: The case of DBpedia. In Extended Semantic Web Conference (pp. 16-27). Springer, Berlin, Heidelberg.

Lexical consistency

Who succeeded to …. ?

- Charles VII the Victorious
- Charles 09 Of France
- Louis VII

Risks on data accuracy



Data accuracy and the challenge of semantic data modelling

Quality indicators and thresholds to assess usability

Based on quality issues in source data, what is the risk generating a “good” item?

Graph	Which part of the statements are accurate?
NELL	74%
YAGO	95%

Brumes et pluies
Ô fins d'automne, hivers, printemps trempés de boue,

Endormeuses sai**z**ons ! je vous aime et vous loue

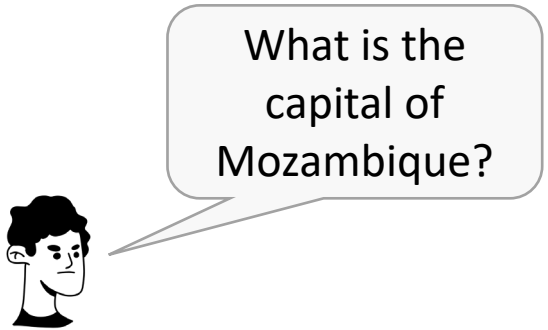
D'envelopper ainsi **n**on coeur et mon cerveau
D'un linceul vaporeux et d'un vague tombeau.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell, “Toward an architecture for never-ending language learning,” in AAAI, 2010

F.-e. M. Suchanek, G. Kasneci, and G. Weikum, “YAGO: A core of semantic knowledge,” in WWW, pp. 697–706, 2007

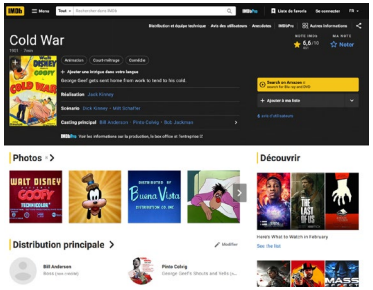
A question-answering system

Answering factual questions

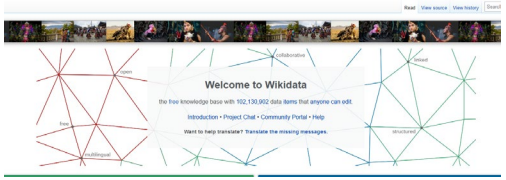


Multiple sources – multiple risks

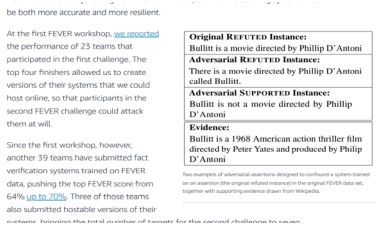
Closed structured data



Open structured data



Unstructured data



be both more accurate and more resilient.

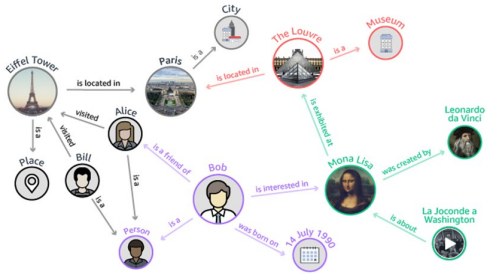
At the first FEVER workshop, we reported the performance of 23 teams that participated in the first challenge. The top four finishers allowed us to create versions of their systems that we could host online, so that participants in the second FEVER challenge could attack them at will.

Since the first workshop, however, another 39 teams have submitted fact verification systems trained on FEVER data, pushing the top FEVER score from 64% up to 70%. Three of those teams also submitted hostable versions of their

Original REFUTED Instance: Bullitt is a movie directed by Phillip D'Antoni
Adversarial REFUTED Instance: There is a movie directed by Phillip D'Antoni called Bullitt.
Adversarial SUPPORTED Instance: Bullitt is not a movie directed by Phillip D'Antoni
Evidence: Bullitt is a 1968 American action thriller film directed by Peter Yates and produced by Philip D'Antoni

Two examples of adversarial assertions designed to confound a system trained on an assertion (the original refuted instance) in the original FEVER data set, together with supporting evidence drawn from Wikipedia.

https://www.amazon.science/blog/the-fever-data-set-what-doesnt-kill-it-will-make-it-stronger

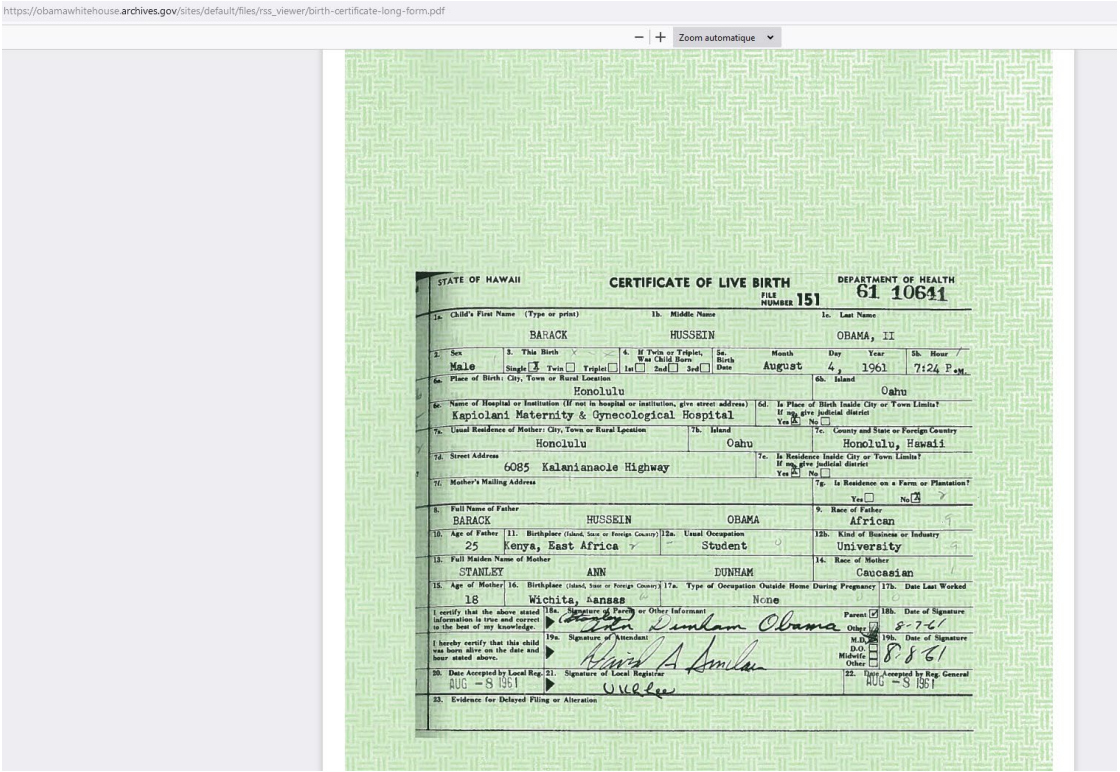


https://aws.amazon.com/fr/neptune/knowledge-graphs-on-aws/

- How do I know what's true?

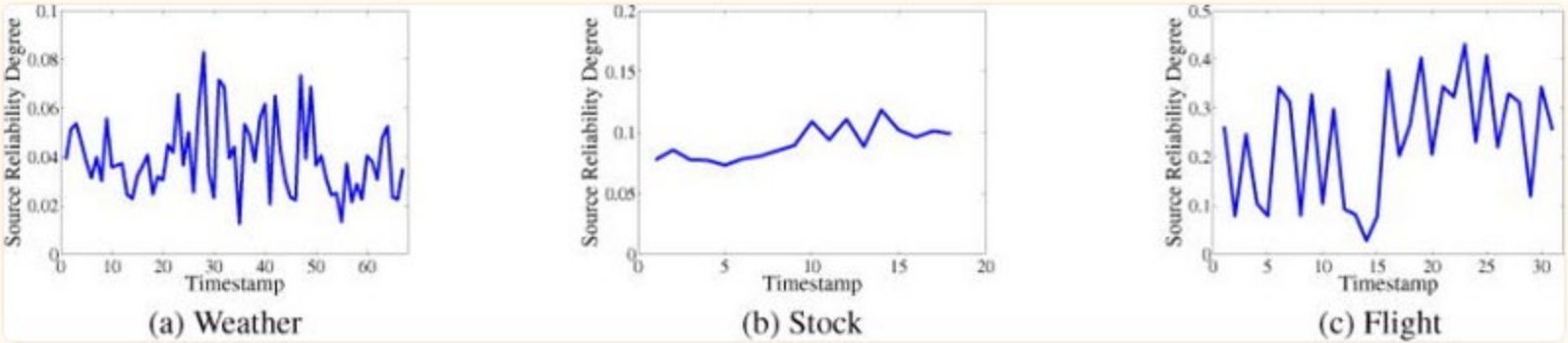
ENDORSE

How many sources agree?



https://obamawhitehouse.archives.gov/sites/default/files/rss_viewer/birth-certificate-long-form.pdf

Data source quality is a moving target



Li, Yaliang, et al. "On the discovery of evolving truth." Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining. 2015.

Consistency with prior knowledge

- Rules:
  - a person cannot manage a company if they are dead
  - a horse cannot fly
- Data:

Table 7: Top 20 male professions in FB3M relative to female using ComplEx embeddings

Profession	Score	C_{male}	$C_{fem.}$
/m/0513qg	0.186	160	8
detective	0.163	27	2
trumpeter	0.161	346	6
gangster	0.146	45	0
private investigator	0.142	18	4
assn. football manager	0.132	587	5
Trombonist	0.131	196	1
session musician	0.130	184	7
sailor	0.119	429	23
bodyguard	0.117	33	2
bandleader	0.115	533	32
assn. football player	0.115	13321	227
samurai	0.114	26	0
music director	0.114	643	29
mastering engineer	0.111	33	1
clergy	0.107	78	4
baseball umpire	0.107	88	0
rabbi	0.105	180	5
Mafioso	0.103	60	0
statistician	0.103	205	3

Table 8: Top 20 female professions in FB3M relative to male using ComplEx embeddings

Profession	Score	$C_{fem.}$	C_{male}
gravure idol	0.210	62	0
fitness professional	0.184	24	12
Nude Glamour Model	0.177	511	1
showgirl	0.171	41	0
nun	0.167	41	0
socialite	0.164	81	11
art model	0.157	22	2
Key Hair Stylist	0.157	43	11
jewellery designer	0.154	39	9
fashion model	0.153	508	32
nurse	0.152	185	20
supermodel	0.151	95	9
Memoirist	0.148	30	35
Adult model	0.147	24	1
pin-up girl	0.146	55	0
dialect coach	0.143	14	8
Prostitute	0.140	63	0
flight attendant	0.137	34	3
ballet dancer	0.135	237	104
Cheerleader	0.133	20	1

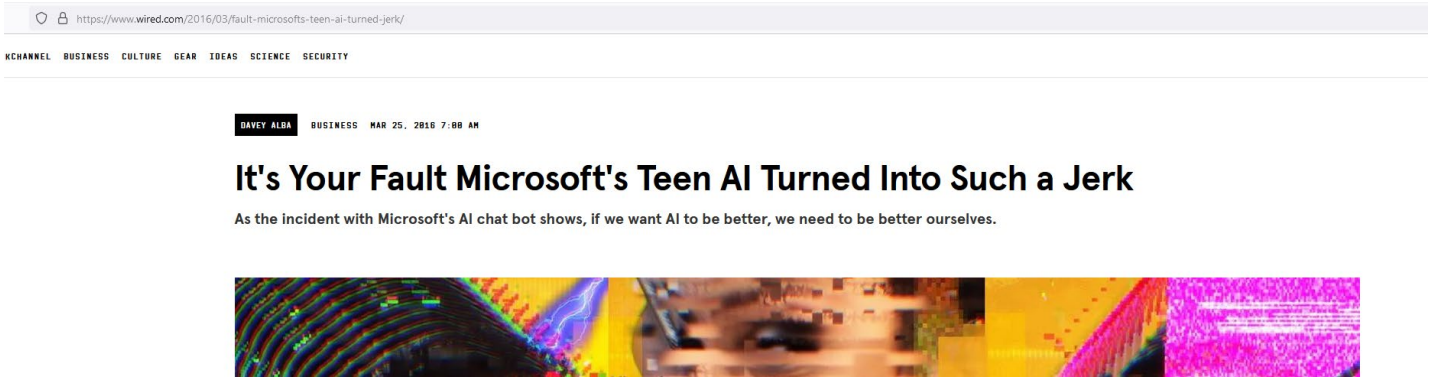
Fisher, J., Palfrey, D., Christodoulopoulos, C., & Mittal, A. (2019). Measuring social bias in knowledge graph embeddings. arXiv preprint arXiv:1912.02761.

https://www.europeana.eu/en/item/03919/public_mistral_joconde_fr_ACTION_CHERCHER_FIELD_1_REF_VALUE_1_50030026886

Labelling, annotating and fixing data manually

- Humans are not perfect

The risk of getting it wrong



https://www.wired.com/2016/03/fault-microsofts-teen-ai-turned-jerk/

KCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

DAVEY ALBA BUSINESS MAR 25, 2016 7:00 AM

It's Your Fault Microsoft's Teen AI Turned Into Such a Jerk

As the incident with Microsoft's AI chat bot shows, if we want AI to be better, we need to be better ourselves.

https://www.wired.com/2016/03/fault-microsofts-teen-ai-turned-jerk/

What’s the truth?

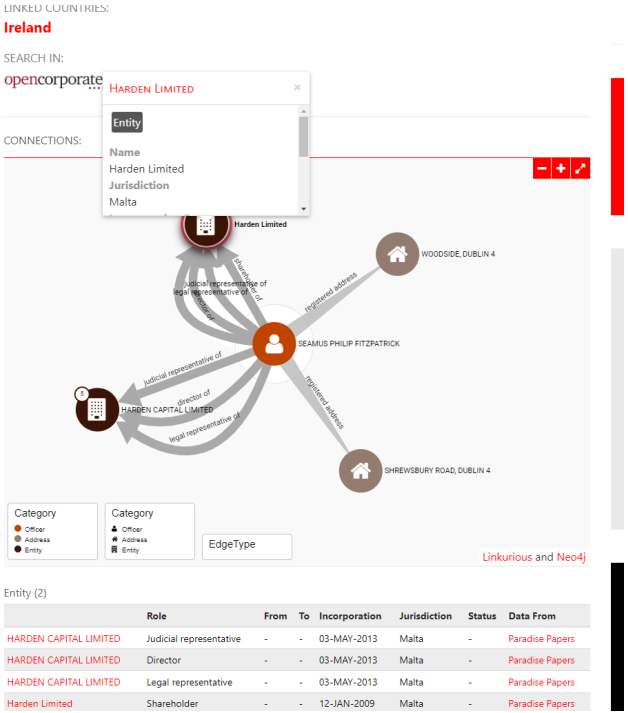
Who is the president of North Korea?

https://commons.wikimedia.org/wiki/File:Mansudae_Grand_Monument_08.JPG

Anti-money laundering

ENDORSE

Finding connections

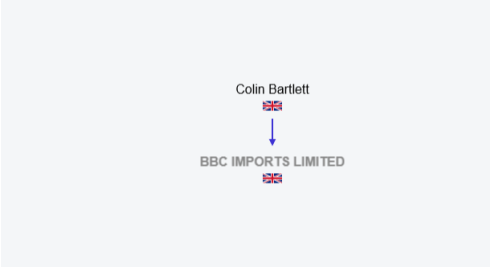


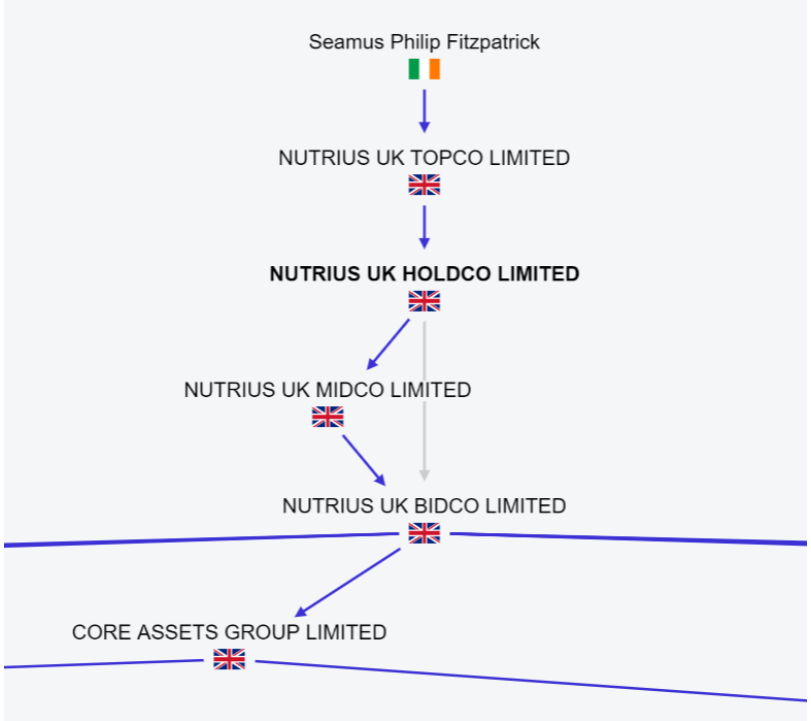
Entity (2)

	Role	From	To	Incorporation	Jurisdiction	Status	Data From
HARDEN CAPITAL LIMITED	Judicial representative	-	-	03-MAY-2013	Malta	-	Paradise Papers
HARDEN CAPITAL LIMITED	Director	-	-	03-MAY-2013	Malta	-	Paradise Papers
HARDEN CAPITAL LIMITED	Legal representative	-	-	03-MAY-2013	Malta	-	Paradise Papers
Harden Limited	Shareholder	-	-	12-JAN-2009	Malta	-	Paradise Papers

https://offshoreleaks.icij.org/nodes/56101617

Finding company structure





https://register.openownership.org/

Entity resolution to establish connections

Is this the same person?

Is this the same company?

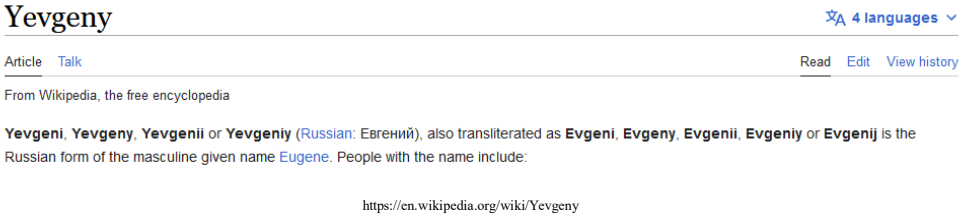
DURAND Jean-Michel	**JM2D**	12 RUE VICTOR MASSÉ 75009 PARIS	504 631 227 00015 RCS PARIS Siège social		kbis
Durand Jean-Michel	**POLE POSITION**	60 RUE MONSIEUR LE PRINCE 75006 PARIS	791 407 091 00010 RCS PARIS Siège social		kbis

https://data.inpi.fr/

ENDORSE

Some issues due to international data collection

- Transliteration
- Multiple names
- Uneven variability of patronyms
- Geographic locations names change
- Geopolitical changes


Yevgeny

4 languages

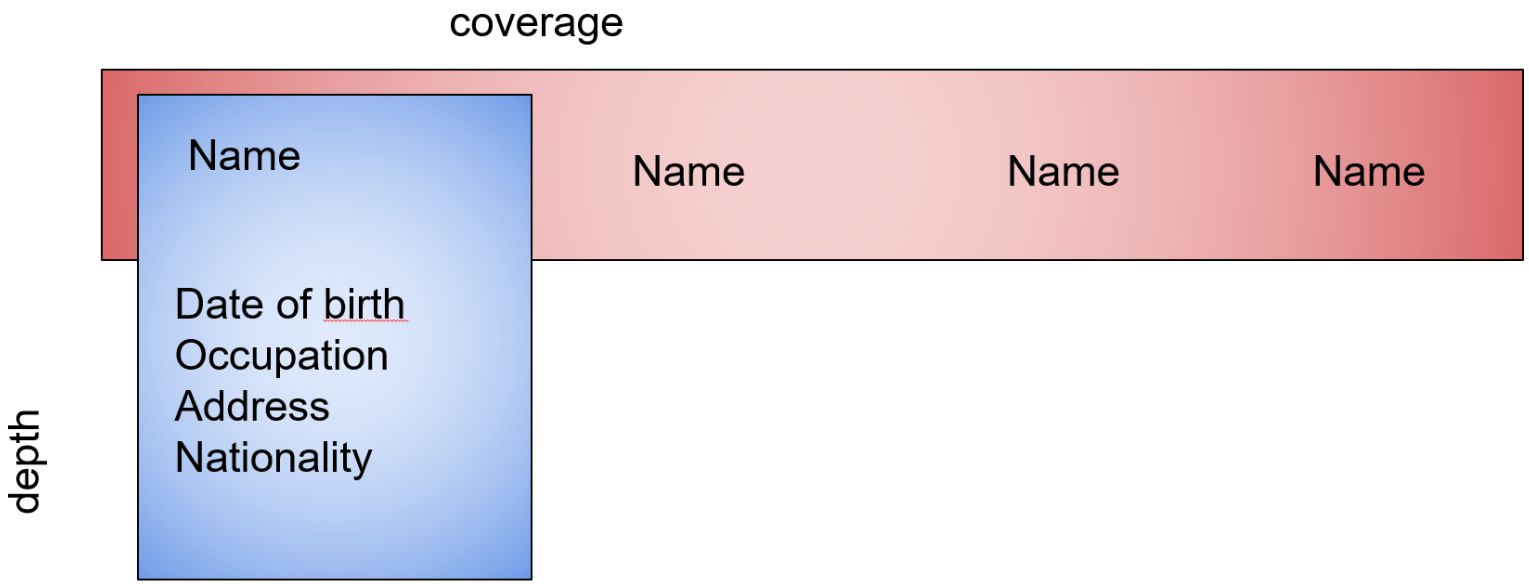
Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

Yevgeni, **Yevgeny**, **Yevgenii** or **Yevgeniy** (Russian: Евгений), also transliterated as **Evgeni**, **Evgeny**, **Evgenii**, **Evgeniy** or **Evgenij** is the Russian form of the masculine given name Eugene. People with the name include:

https://en.wikipedia.org/wiki/Yevgeny

Completeness issues



Sparse data makes it difficult to answer:
‘is this the same person?’

Assess a risk on data freshness

"Experts say 2 percent of records in a customer file become obsolete in one month because customers die, divorce, marry, and move."

Eckerson, W. W. (2002). Data quality and the bottom line: Achieving business success through a commitment to high quality data. The Data Warehousing Institute, 1-36.

https://data.inpi.fr/entreprises/450516737?q=wimi#450516737

- The risk of data obsolescence

ENDORSE

Identité

Dénomination
SCI WIMI

SIREN (siège)
450 516 737

N° de gestion
2003D00300

Début d'activité
12/09/2003

Durée de la personne morale
99 ans

Date de clôture
31 Décembre

Forme juridique
Société civile

Activité principale
Acquisition, administration, gestion par location ou autrement de tous immeubles et plus spécialement, acquisition sur la commune de Vic-Fezensac d'une grange avec terrain attenant, le tout sis chemin de l'Abattoir.

Capital social
1 600.00 €

Adresse du siège
chemin de l'Abattoir 32190 Vic-Fezensac FRANCE

Département du siège
32

Représentants

Pour plus d'informations sur les représentants, veuillez **vous connecter**

Nom, Prénom(s)
LAMBERT Remi, Robert, Paul
(Gérant, Associé indéfiniment responsable)

Date de naissance (mm/aaaa)
06/1966

Nom, Prénom(s)
COCHONNEAU William, Louis, Marcel
(Associé indéfiniment responsable)

Date de naissance (mm/aaaa)
04/1956

Établissements

Type d'établissement
Siège et principal

Début d'activité
12/09/2003

Origine du fonds
Création

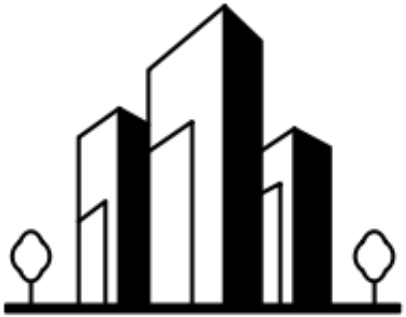
Type d'exploitation
Exploitation directe

Activité
Acquisition, administration, gestion par location ou autrement de tous immeubles et plus spécialement, acquisition sur la commune de Vic-Fezensac d'une grange avec terrain attenant, le tout sis chemin de l'Abattoir.

Adresse
chemin de l'Abattoir 32190 Vic-Fezensac FRANCE

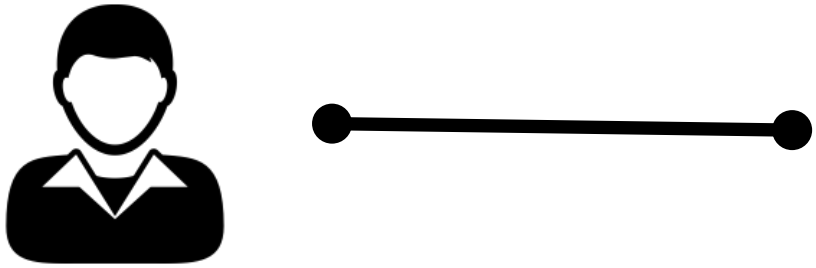
Exploration of the UK company registry

UK Companies House



ENDORSE

Exploration of the UK company registry



Search for a company or officer

Advanced company search

MINISTRY OF FINANCE OF SAUDI ARABIA LTD

Company number **10625481**

Follow this company | File for this company

Overview | Filing history | People | More

Registered office address
27 Old Gloucester Street, London, United Kingdom, WC1N 3AX

Company status
Active

Company type
Private limited Company

Incorporated on
17 February 2017

ENDORSE

Some financial data on the company

MINISTRY OF FINANCE OF SAUDI ARABIA LTD **Registered Number 10625481**

Balance Sheet as at 28 February 2022

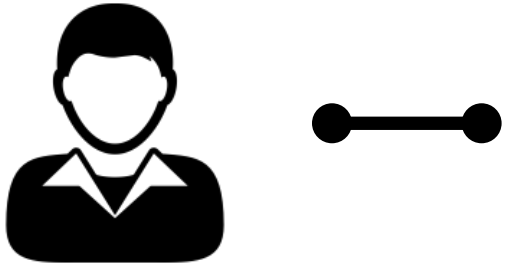
	2022	2021
	£	£
Called up share capital not paid	1000000000	1000000000
Net assets	**1000000000**	**1000000000**
Issued share capital		
10000000 Ordinary Shares of £ 100 each	1000000000	1000000000
Total Shareholder funds	**1000000000**	**1000000000**

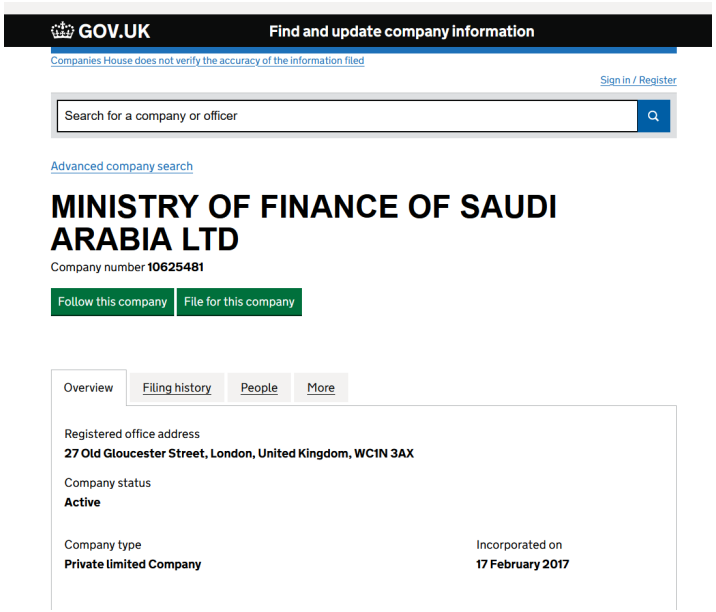
STATEMENTS

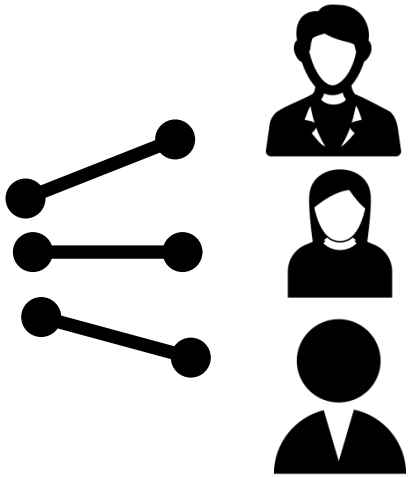
https://find-and-update.company-information.service.gov.uk/company/10625481/filing-history

ENDORSE

Business associates







ENDORSE

Indirect connections to other companies and business associates

BANQUE DU CANADA LTD

Company number **12342074**

Follow this company | File for this company

Overview | Filing history | People | More

Officers | Persons wit

Filter officers

☐ Current officers

BANQUE CANTONALE DE GENÈVE LTD

Company number **09931624**

Follow this company | File for this company

Overview | Filing history | People | More

Registered office address
27 Old Gloucester Street, London, WC1N 3AX

Company status
Active

Company type
Private limited Company

Incorp
29 De

Advanced company search

MINISTRY OF FINANCE OF THE STATE OF QATAR LTD

Company number **11656934**

Follow this company | File for this company

Overview | Filing history | People | More

Registered office address
27 Old Gloucester Street, London, United Kingdom, WC1N 3AX

Balance Sheet as at 28 December 2021

	2021	2020
	£	£
Called up share capital not paid	1000000000	1000000000
Net assets	**1000000000**	**1000000000**
Issued share capital		
10000000 Ordinary Shares of £ 100 each	1000000000	1000000000
Total Shareholder funds	**1000000000**	**1000000000**

ENDORSE

Data collection method

GOV.UK Find and update company info

Companies House does not verify the accuracy of the information filed

Search for a company or officer

Advanced company search

UKCH LIMITED

Company number **10115452**

Follow this company | File for this company

https://www.theguardian.com/business/2022/nov/08/companies-house-is-dysfunctional-and-facilitating-mps-told

Mail | YouTube | Maps | Gartic Phone - Le jeu ... | Jouez à notre Escape ... | Gartic Phone - Le jeu ... | owa.uni...

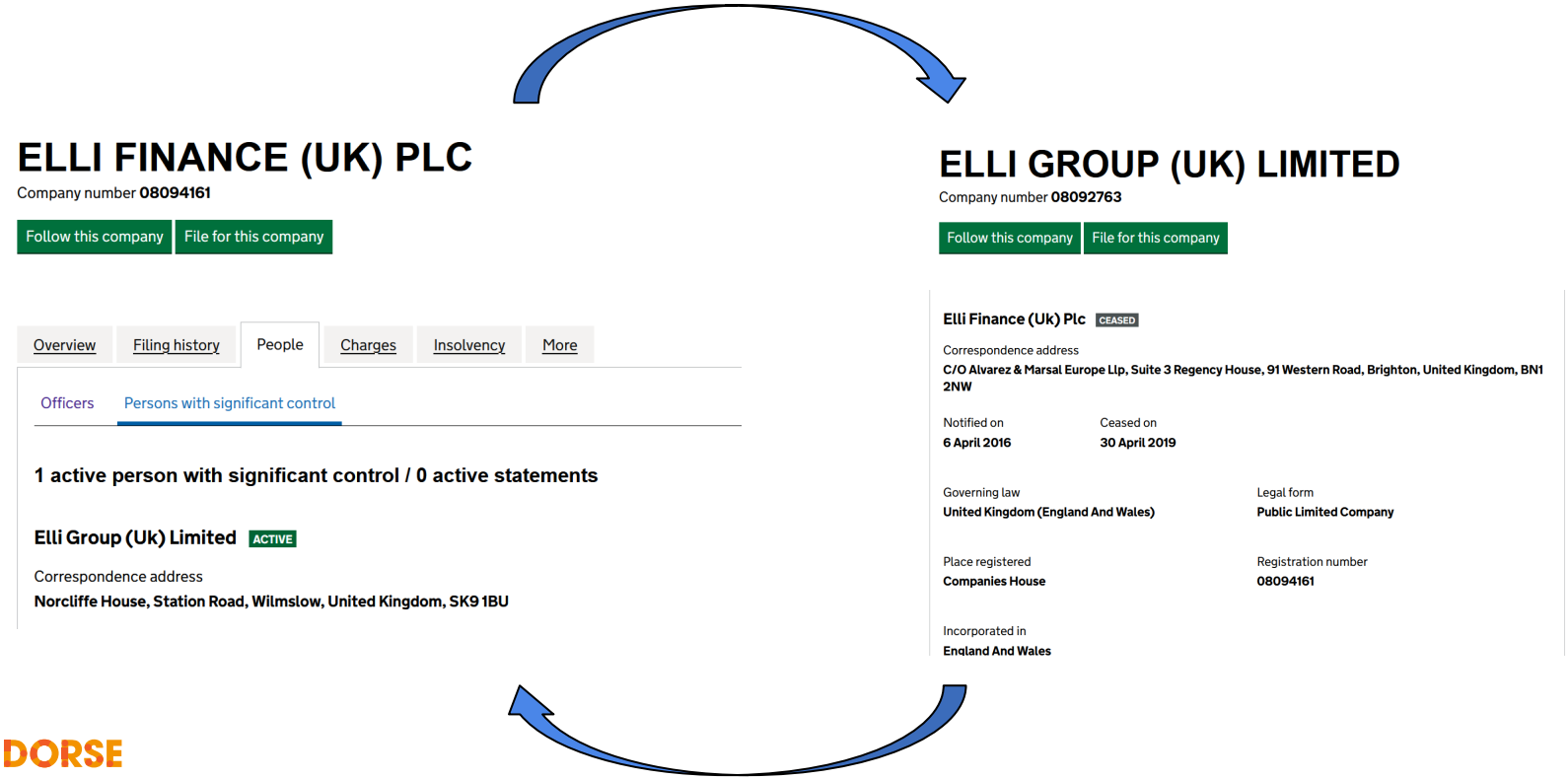
Companies House is dysfunctional and facilitating fraud, MPs told

Less verification for someone to set up fraudulent shell firm than to borrow a library book, risk managers say

Anti-fraud bosses at NatWest and HSBC have criticised the online register of UK-based companies. Photograph: Dominic Lipinski/PA

The anti-fraud leader at the trade body UK Finance has said the government needs to fix the "dysfunctional" Companies House because it is helping to facilitate business fraud.

Data input interfaces and honest(?) mistakes



ENDORSE

Lack of data standardizations makes connections difficult

London Laundromat: Police seize £2 million profits of Italian mafia gang held in British banks

this company

- 29 Chichele Road, London, United Kingdom, NW2 3AN
- Office 2092, No.1, Fore Street, London, England, EC2Y 5EJ
- 20-22, Wenlock Road, London, England, N1 7GU
- ….

Overview | Filing history | People | More

Registered office address
7 29 Chichele Road, London, NW2 3AN

Company status
Active

Company type
Private limited Company

Incorporated on
14 October 2010

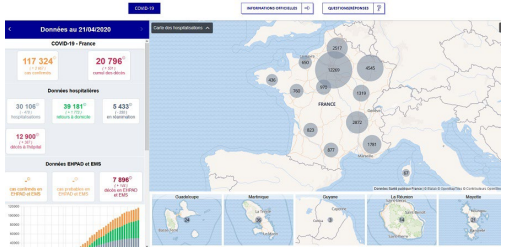
ENDORSE

https://www.standard.co.uk/news/crime/ps2million-mafia-cash-laundered-in-london-is-seized-by-police-a4506351.html

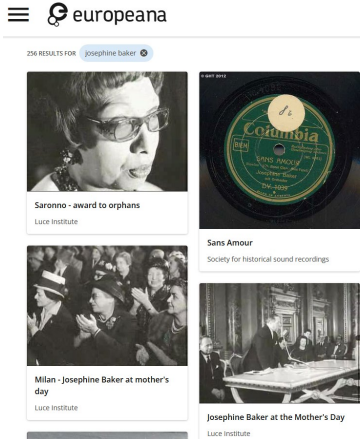
Aim for excellence but build services that are resilient to imperfections

Data has multiple roles

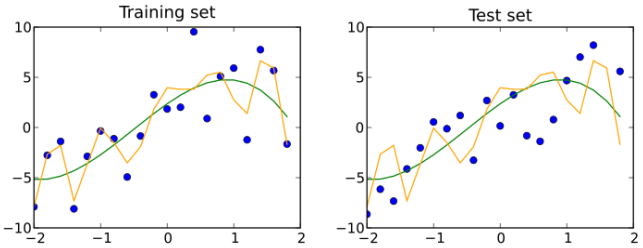
Represent data



Provide access in a new context

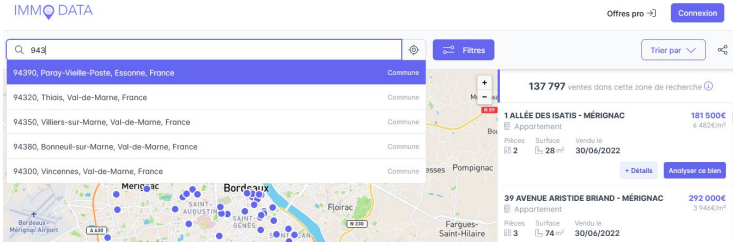


Train model

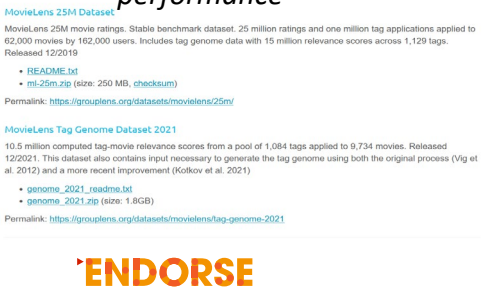


Skbkekas, CC BY 3.0 <https://creativecommons.org/licenses/by/3.0>, via Wikimedia Commons

Validate own data



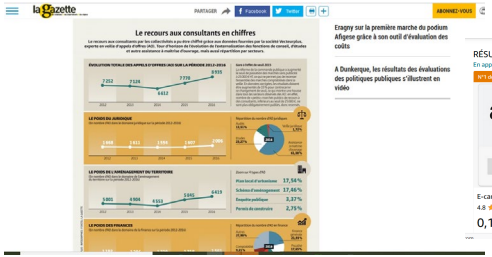
Validate algorithm performance



ENDORSE

Enrich own data

Analyze data to support (political) decision making



Personalize experience

Data quality as fitness for use or fitness for purpose

- A multiplicity of reuse contexts
- The data has a multiplicity of roles

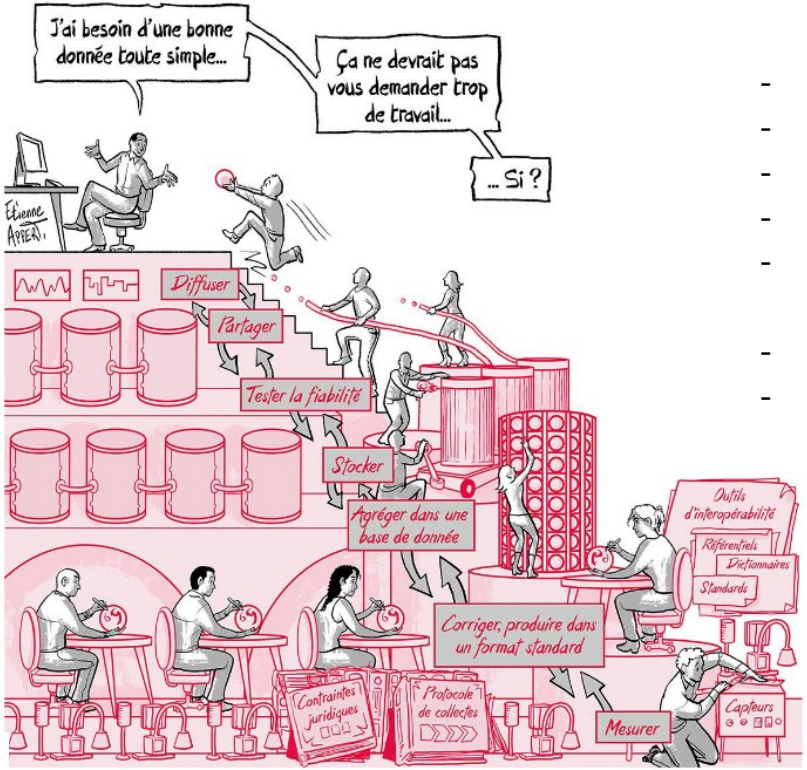
not all are well known in open environments

Quality is a construction

- I need good data, it's very simple

… that should not require a lot of work

… or does it?



- *Distribute*
- *Share*
- *Assess quality*
- *Store*
- *Aggregate in a database*
- *Fix, standardize*
- *Measure*

GOUVERNEMENT
Ministère de la Transition écologique et de la Cohésion des territoires
Ministère de la Transition énergétique

https://www.ecologie.gouv.fr/sites/default/files/Principes_generaux_qualite_des_donnees_MTE_MTECT_1-1.pdf

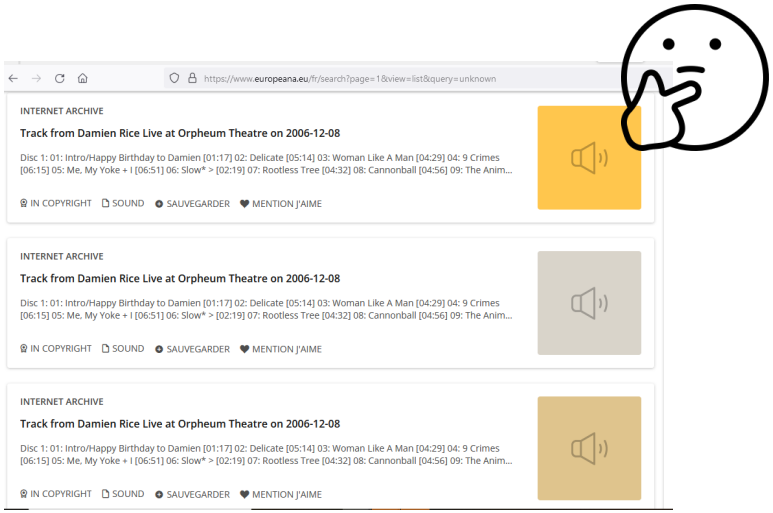
Illustration Etienne Appert

ENDORSE

- Explain it

Show impact

Most of the data is not bad in its original environment


https://www.europeana.eu/fr/search?page=1&view=list&query=unknown

INTERNET ARCHIVE

Track from Damien Rice Live at Orpheum Theatre on 2006-12-08

Disc 1: 01: Intro/Happy Birthday to Damien [01:17] 02: Delicate [05:14] 03: Woman Like A Man [04:29] 04: 9 Crimes [06:15] 05: Me, My Yoke + I [06:51] 06: Slow* > [02:19] 07: Rootless Tree [04:32] 08: Cannonball [04:56] 09: The Anim...

INTERNET ARCHIVE

Track from Damien Rice Live at Orpheum Theatre on 2006-12-08

Disc 1: 01: Intro/Happy Birthday to Damien [01:17] 02: Delicate [05:14] 03: Woman Like A Man [04:29] 04: 9 Crimes [06:15] 05: Me, My Yoke + I [06:51] 06: Slow* > [02:19] 07: Rootless Tree [04:32] 08: Cannonball [04:56] 09: The Anim...

INTERNET ARCHIVE

Track from Damien Rice Live at Orpheum Theatre on 2006-12-08

Disc 1: 01: Intro/Happy Birthday to Damien [01:17] 02: Delicate [05:14] 03: Woman Like A Man [04:29] 04: 9 Crimes [06:15] 05: Me, My Yoke + I [06:51] 06: Slow* > [02:19] 07: Rootless Tree [04:32] 08: Cannonball [04:56] 09: The Anim...

Coronavirus: Boris Johnson unable to say how many people weren't traced due to 16,000 missed cases

Acknowledge excellence is a target

- But understand the risks
- To some extent it is possible to account for imperfection and work around it

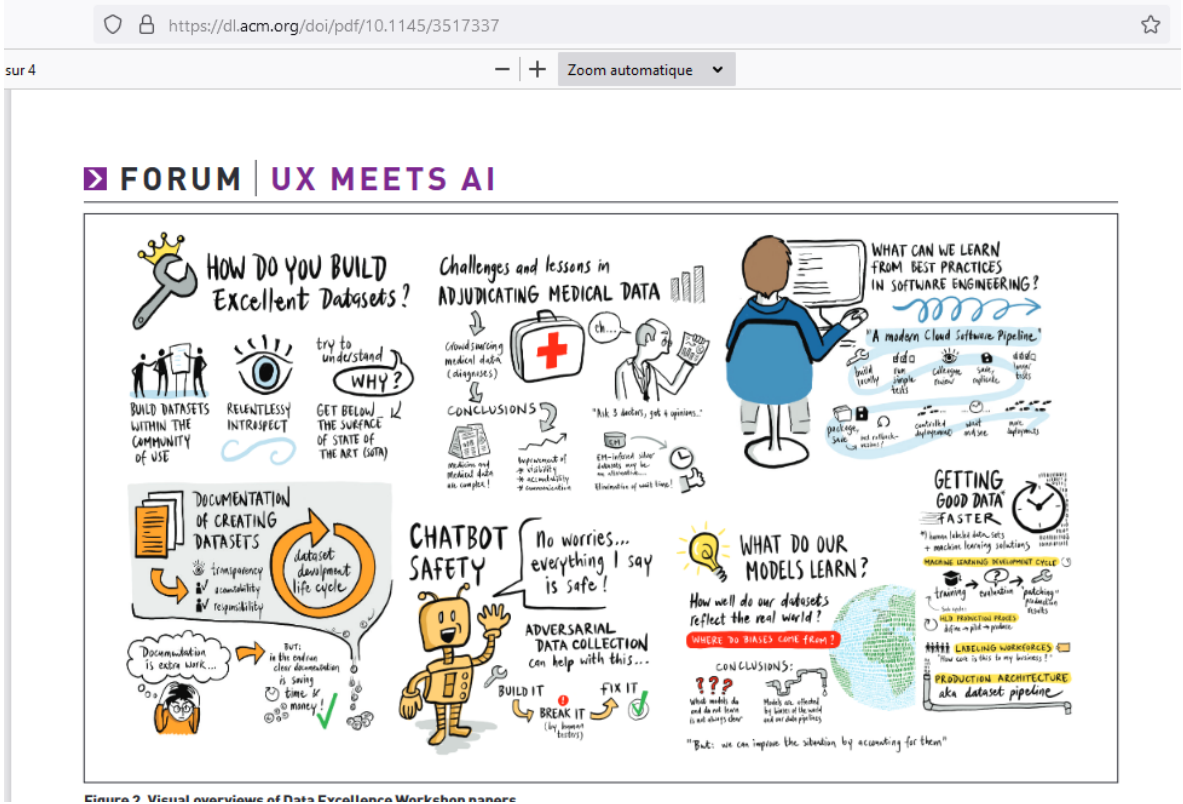


Figure 2. Visual overviews of Data Excellence Workshop papers

Aroyo, L., Lease, M., Paritosh, P., & Schaekermann, M. (2022). Data excellence for AI: why should you care?. Interactions, 29(2), 66-69.

Quality, excellence and perfectionism

“Osborn employs the concept of perfectionism to describe a hyper-emphasis on exactitude and precision-in this case, quality gone awry by being taken to an extreme.”

Thomas, Sarah E. "Quality in bibliographic control." (1996).
https://www.ideals.illinois.edu/items/7998/bitstreams/27643/stream

Muriel Foulonneau