Data Interoperability and Semantics

< Part 4. The value of data >

Data Interoperability and Semantics
Outline

Data Interoperability and Semantics

Part 4. The value of data

Part 4.1. Value as one of the V's of Big Data

Orders of magnitude of the total amount of data created, captured, copied, and consumed globally ?

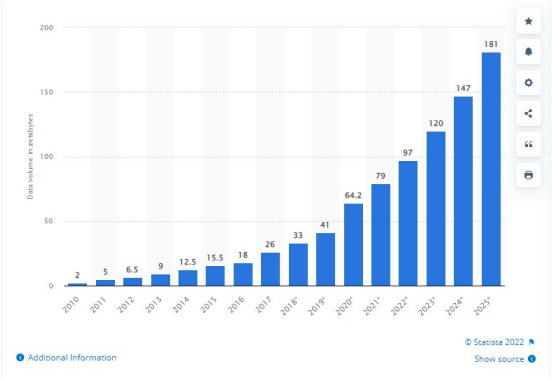
Multiple-byte units V·T·E

Decimal			Binary		
Value	Metric		Value	IEC	Legacy
1000	kB	kilobyte	1024	KiB kibibyte	KB kilobyte
1000^2	MB	megabyte	1024^2	MiB mebibyte	MB megabyte
1000^3	GB	gigabyte	1024^3	GiB gibibyte	GB gigabyte
1000^4	TB	terabyte	1024^4	TiB tebibyte	TB terabyte
1000^5	PB	petabyte	1024^5	PiB pebibyte	–
1000^6	EB	exabyte	1024^6	EiB exbibyte	–
1000^7	ZB	zettabyte	1024^7	ZiB zebibyte	–
1000^8	YB	yottabyte	1024^8	YiB yobibyte	–

Orders of magnitude of data

Projected figures 2025 for the total amount of data created, captured, copied, and consumed globally

530%
increase of global data volume
From 33 zettabytes in 2018 to 175 zettabytes



https://www.statista.com/statistics/871513/worldwide-data-created/

https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#projected-figures-2025

Projected figures 2025

530%
increase of global data volume
From 33 zettabytes in 2018 to 175 zettabytes

€829 billion
value of data economy in the EU27
From €301 billion (2.4% of EU GDP) in 2018

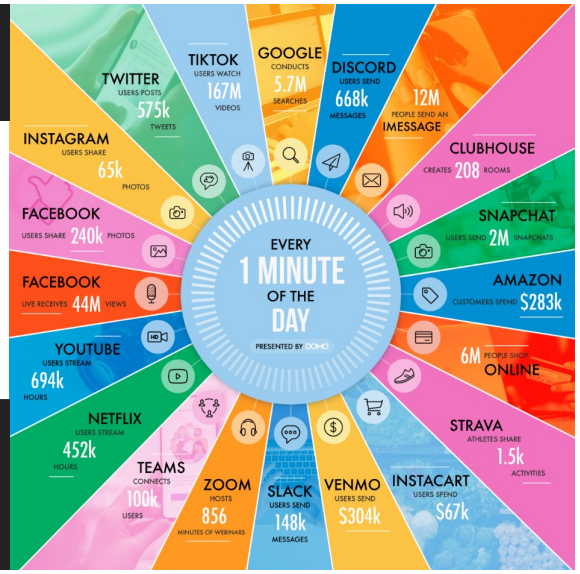
10.9 million
data professionals in the EU27
From 5.7 million in 2018

65%
Percentage of EU population with basic digital skills
From 57% in 2018

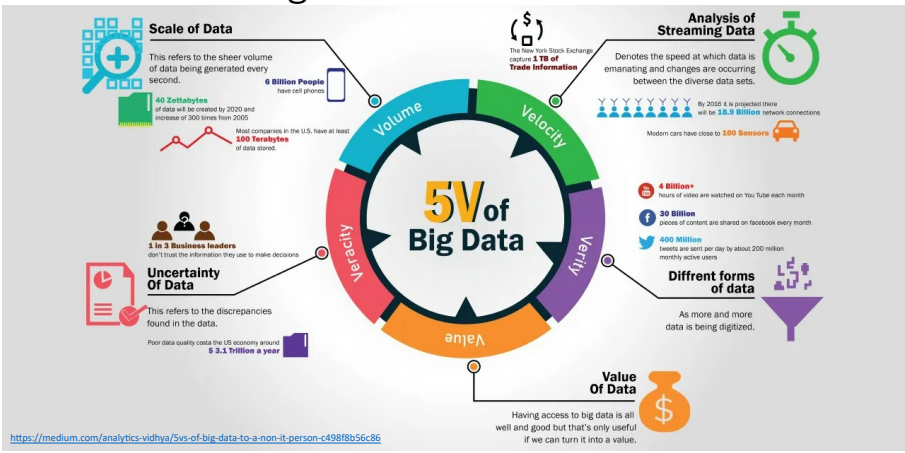
https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#projected-figures-2025

This happens every minute of every day, according to DOMO
(Source: Data Never Sleeps 9.0)

https://www.domo.com/learn/infographic/data-never-sleeps-9



The 5 V's of Big Data ?



https://medium.com/analytics-vidhya/5vs-of-big-data-to-a-non-it-person-c498f8b56c86

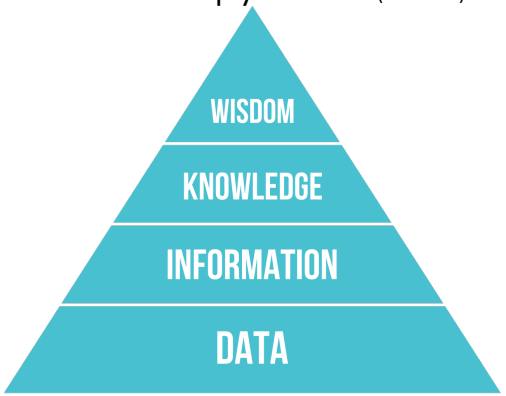
Data Interoperability and Semantics

Part 4. The value of data

Part 4.2. Data, Information, Knowledge, Metadata, …

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

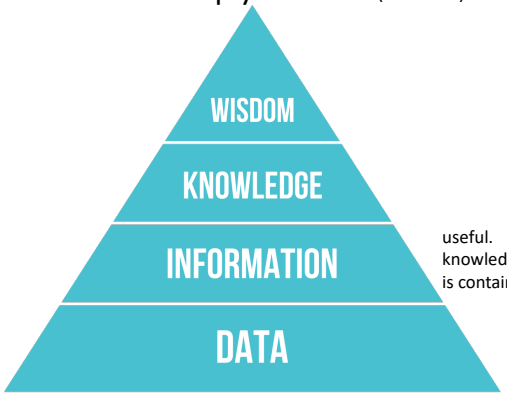
The DIKW pyramid (~1982)



useless.
symbols or signs, representing stimuli or signals, that are "of no use until...in a usable (that is, relevant) form

https://en.wikipedia.org/wiki/DIKW_pyramid , and chosen definitions

The DIKW pyramid (~1982)

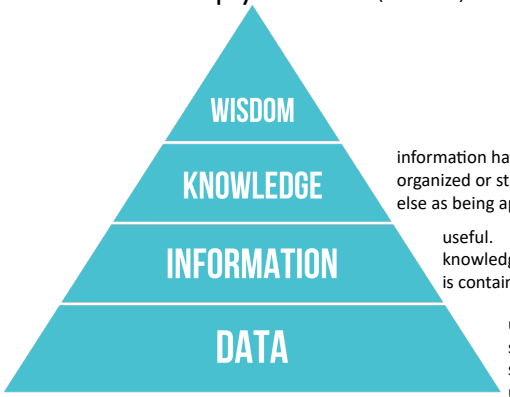


useful.
knowledge by description ("information is contained in descriptions")

useless.
symbols or signs, representing stimuli or signals, that are "of no use until...in a usable (that is, relevant) form

https://en.wikipedia.org/wiki/DIKW_pyramid , and chosen definitions

The DIKW pyramid (~1982)



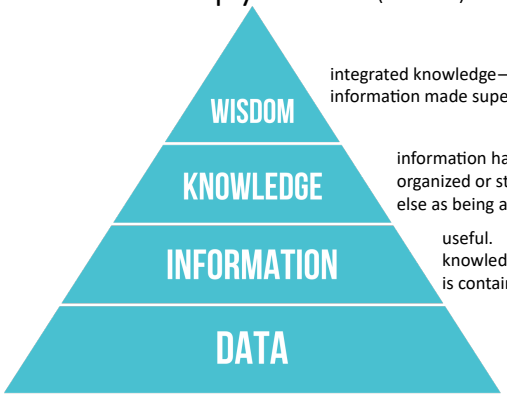
information having been processed, organized or structured in some way, or else as being applied or put into action

useful.
knowledge by description ("information is contained in descriptions")

useless.
symbols or signs, representing stimuli or signals, that are "of no use until...in a usable (that is, relevant) form

https://en.wikipedia.org/wiki/DIKW_pyramid , and chosen definitions

The DIKW pyramid (~1982)



integrated knowledge—
information made super-useful

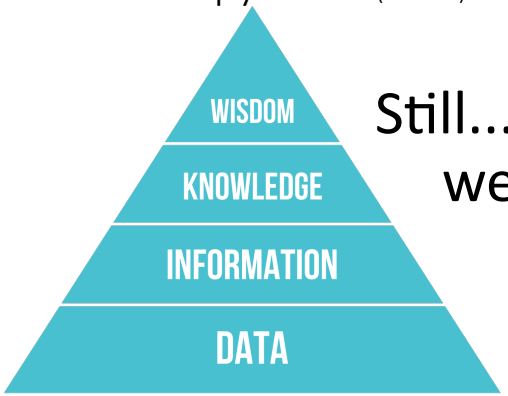
information having been processed, organized or structured in some way, or else as being applied or put into action

useful.
knowledge by description ("information is contained in descriptions")

useless.
symbols or signs, representing stimuli or signals, that are "of no use until...in a usable (that is, relevant) form

https://en.wikipedia.org/wiki/DIKW_pyramid , and chosen definitions

The DIKW pyramid (~1982)



Still...
we often use
« Data »
to generalize

Definitions on Data

— ISO/IEC20546:2019 (Big data – Overview and vocabulary)

Dataset
Identifiable collection of data available for access or download in one or more formats

Data
Re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing
Note 1 to entry: Data can be processed by humans or by automatic means.

Metadata
Data about data or data elements, possibly including their data descriptions and data about data ownership, access paths, access rights and data volatility

Information
Data that are processed, organised and correlated to produce meaning.
Note 1 to entry: Information concerns facts, concepts, objects, events, ideas, processes, etc.

Metadata

"data that provides information about other data"

- **Descriptive metadata** – the descriptive information about a resource.
  - For discovery and identification.
  - Ex: title, abstract, author, keywords.
- **Structural metadata** – containers of data, how compound objects are put together
  - Ex, how pages are ordered to form chapters.
  - Ex: types, versions, relationships, and other characteristics of digital materials.
- **Administrative metadata** – the information to help manage a resource
  - Ex: resource type, permissions, time, when and how it was created.
- **Reference metadata** – contents and quality
  - For quality assessment of the data
  - Ex: conceptual metadata, quality metadata, methodological metadata.
- **Statistical metadata**, also called process data,
  - May describe processes that collect, process, or produce statistical data.
  - Number of rows, columns, etc.
- **Legal metadata** –
  - Ex: license, creator, copyright holder

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

Data Interoperability and Semantics

Part 4. The value of data

Part 4.3. Interoperability unlocks the value of data

Data Analytics vs Data Processing

— ISO/IEC20546:2019 (Big data – Overview and vocabulary)

Data Analytics
Composite concept consisting of data acquisition, data collection, data validation, data processing, including data quantification, data visualisation and data interpretation

Data Processing
Systematic performance of operations upon data
Note 1 to entry: Example: Arithmetic or logic operations upon data, merging or sorting of data, or operations on text, such as editing, sorting, merging, storing, retrieving, displaying, or printing.

Data silos problem illustrated

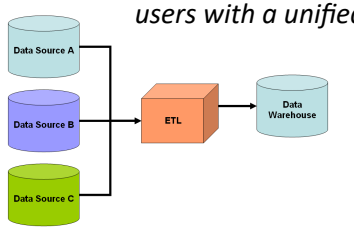
Repositories of fixed data that are isolated, incompatible, or not integrated

Example of silos https://factory.dev/pimcore-knowledge-base/whitepaper/data-silos

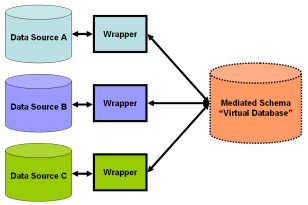
Data integration

*combining **heterogeneous data** and providing users with a unified view of them.*

Simple schematic for a data warehouse.
The Extract, transform, load (ETL) process extracts information from the source databases, transforms it and then loads it into the data warehouse.



Simple schematic for a data-integration solution.
A system designer constructs a mediated schema against which users can run queries. The virtual database interfaces with the source databases via wrapper code if required.



Sub-areas:
- Data warehousing
- Data migration
- Enterprise application/information integration
- Master data management

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

Interoperability vs Portability

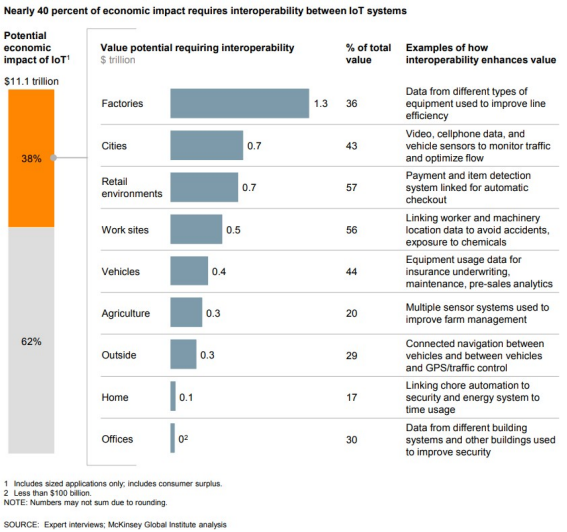
— IEEE Standard Computer Dictionary
— ISO/IEC19941:2017 (Cloud computing – interoperability and portability)

Interoperability
Ability of two or more systems or components to exchange information and to mutually use the information that has been exchanged

Portability
Ability to easily transfer data from one system to another without being required to re-enter data

Interoperability as an enabler for the potential value of data

ex:
Internet of Things



https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world

Interoperability at different levels: from the hardware to policies

- **Policies**: Laws, regulations, norms, ...
- **Behaviour**: Processus, services, operations, ...
- **Semantic**: Knowledge domain vocabularies, taxonomies, ontologies, ...
- **Syntactic**: Encoding, Formats, Schemas, ...
- **Transport:** see also Open Systems Interconnections (OSI) network layers
- **OS:** Kernel, File systems,...
- **Hardware**: Instruction Set Architectures, ...

ISO/IEC definitions of interoperability

Transport interoperability
interoperability where information exchange uses an established communication infrastructure between the participating systems
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Syntactic interoperability
interoperability such that the formats of the exchanged information can be understood by the participating systems
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Semantic data interoperability
interoperability so that the meaning of the data model within the context of a subject area is understood by the participating systems
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Behavioural interoperability
interoperability so that the actual result of the exchange achieves the expected outcome
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Policy interoperability
interoperability while complying with the legal, organizational, and policy frameworks applicable to the participating systems
— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Data Interoperability – a definition

Data interoperability addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data.

— https://datainteroperability.org/

Syntactic interoperability
interoperability such that the formats of the exchanged information can be understood by the participating systems

— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Semantic data interoperability
interoperability so that the meaning of the data model within the context of a subject area is understood by the participating systems

— ISO/IEC 22123-1:2021, Cloud computing — Part 4: Vocabulary

Norm-based interoperability

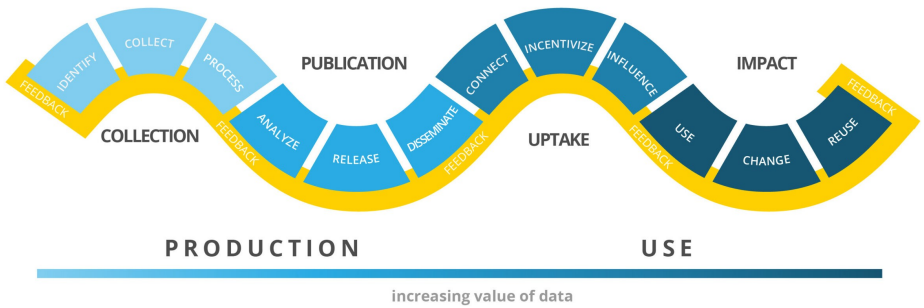
- Hardware and/or software standards
- Standard = Detailed set of technical requirements intended to establish a certain uniformity (in a field of hardware or software development)
- Standard ---- Norm: no clear boundary
  - docx: Microsoft
  - HTML, XML: World Wide Web consortium (W3C)
  - SIM cards: European Telecommunication Standard Institute (ETSI)
  - Postscript became standardized after release
  - De facto standards (not produced by a standard development organization). example: csv, json, rar, pdf, java, flash

Data Interoperability and Semantics

Part 4. The value of data

Part 4.4. Sharing data generates economic and societal value

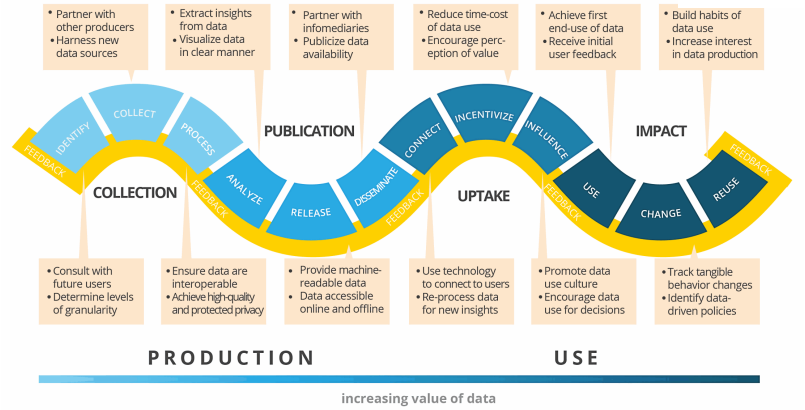
The Data Value chain



Open Data Watch. "The data value chain: Moving from production to impact." *Data2X.* *https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact* (2018).

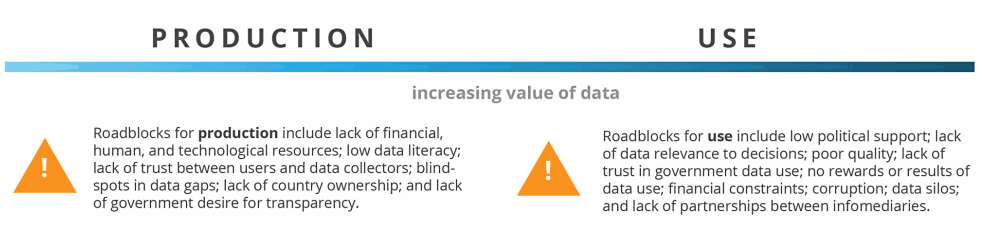
MARKERS — Potential achievements within each process of the value chain mark progress towards data impacts.

The Data Value chain



Open Data Watch. "The data value chain: Moving from production to impact." *Data2X*. https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact (2018).

The Data Value roadblocks



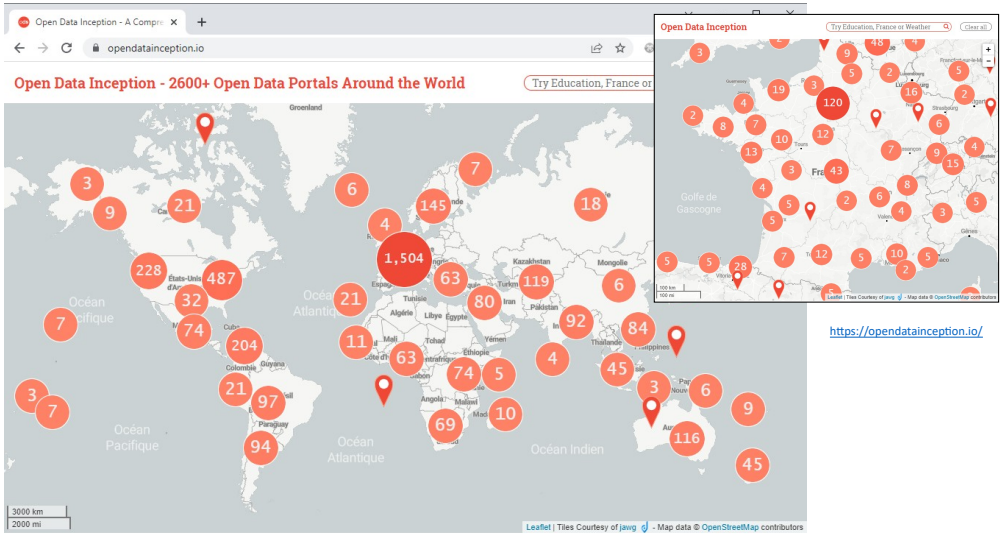
Open Data Watch. "The data value chain: Moving from production to impact." *Data2X*. https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact (2018).

Open Data in the US

- Data.gov 2009
- Legal framework:
  - The U.S. Open Government Directive of December 8, 2009, required that all agencies post **at least three high-value data sets online** and register them on Data.gov **within 45 days**
  - OPEN Government Data Act, as part of the Foundations for Evidence Based Policymaking Act (2019)

Open Data in France

- France at the forefront of Open Data in Europe:
  - Légifrance 1999
- Legal framework:
  - "The society has the right of requesting account from any public agent of its administration." (Declaration of rights of man and of the citizen of 1789)
  - Law on the liberty of access to administrative documents (1978)
  - Euopean directove 2003 + French Law 2005 + Decree 2011
  - Bill on a Digital Republic (2016)
  - The law on Energy Transition (2015)
- 2014: Chief Data Officer in the French public administration



https://opendatainception.io/

Open Data in Europe

EUROPEAN DATA PORTAL

The European Data Portal:
Opening up Europe's public data

data.europa.eu
The official portal for European data

Open Data, Open Content, and Open Knowledge

The Open Definition

The Open Definition sets out principles that define "openness" in relation to **data and content**.

It makes **precise** the meaning of "open" in the terms **"open data"** and **"open content"** and thereby ensures **quality** and encourages **compatibility** between different pools of open material.

It can be summed up in the statement that:

> "Open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness)."

Put most succinctly:

> "Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**"

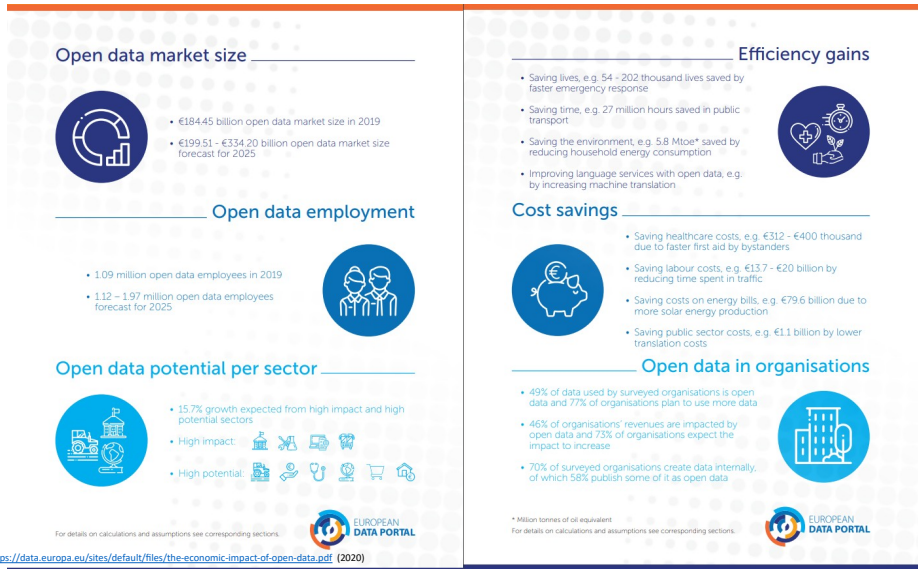
http://opendefinition.org/

Open Work

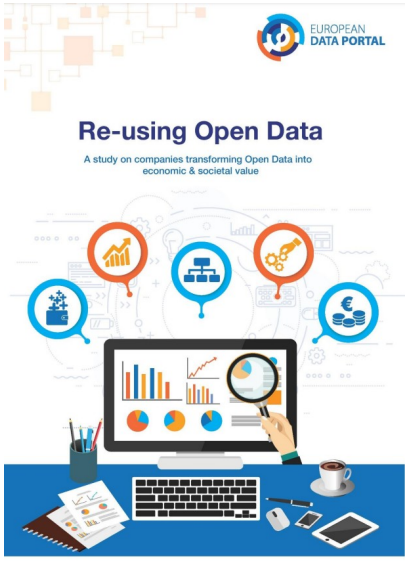
1. Open License or Status
   The **work** *must* be in the **public domain** or provided under an open **license**
2. Access
   The **work** *must* be provided as a whole and at no more than a reasonable one-time reproduction cost, and *should* be downloadable via the Internet without charge.
3. Machine Readability
   The **work** *must* be provided in a form readily processable by a computer and where the individual elements of the work can be easily accessed and modified.
4. Open Format
   The **work** *must* be provided in an open format.

http://opendefinition.org/

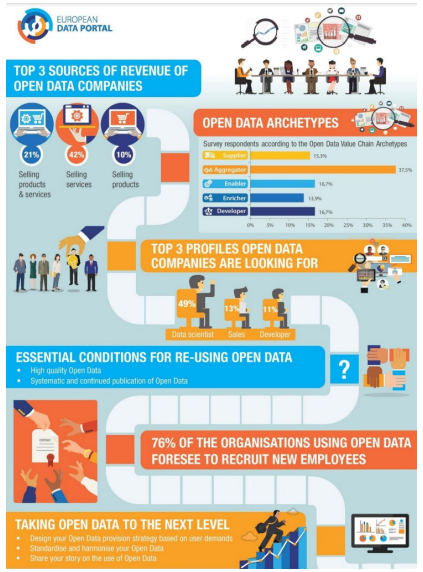
Economic impact of Open Data



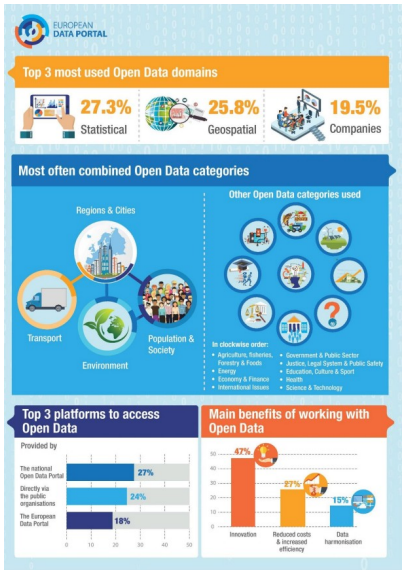
https://data.europa.eu/sites/default/files/the-economic-impact-of-open-data.pdf (2020)



https://data.europa.eu/sites/default/files/re-using_open_data.pdf (2020)



https://data.europa.eu/sites/default/files/re-using_open_data.pdf (2020)



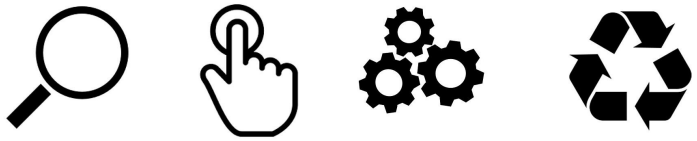
Data Interoperability and Semantics

Part 4. The value of data

Part 4.5. Machine-actionability of data increases its value

FAIR Principles

Findable **A**ccessible **I**nteroperable **R**eusable



The FAIR principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

https://www.go-fair.org/fair-principles/

FAIR Guiding Principles for scientific data management and stewardship (2016)

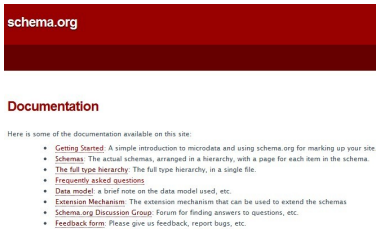
Source : Australian National Data Service (ANDS)

https://www.go-fair.org/fair-principles/

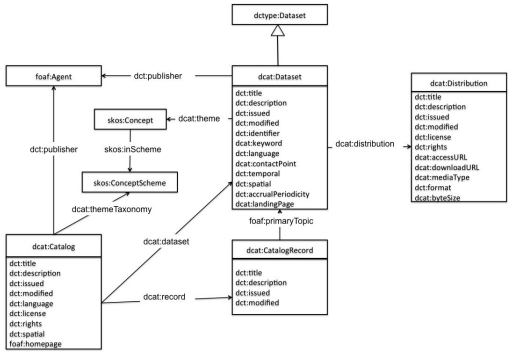
FAIR Guiding Principles for scientific data management and stewardship (2016)

Example of standards for Metadata

Dublin Core



schema.org



Data Catalog Vocabulary (DCAT) - Version 2
W3C Recommendation 04 February 2020

- Demo Dublin core or schema.org in web pages
- Demo DCAT for open data portal catalogs

Data Interoperability and Semantics

< Part 4. The value of data >

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data