15 principles in 20 min

Carlos Utrilla Guerrero


FAIR

FAIR explained in 2 min

https://www.youtube.com/watch?v=5OeCrQE3HhE&t=5s

Challenge:

Understand 15 principles in 20 min

My suggestions, must-read:
1- The FAIR Guiding Principles for scientific data management and stewardship 2016
2- FAIR Principles Interpretations and Implementation Considerations, 2020
3- Cloudy, increasingly FAIR: revising the FAIR Data Guiding Principles for the EOSC 2017

Importantly, it is our intent that **the principles apply not only to 'data'** in the conventional sense, **but also to the algorithms, tools, and workflows** that led to that data.

All scholarly digital research objects—from data to analytical pipelines—**benefit from** application of **these principles**, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

Wilkinson et al. 2016

But today, our main focus is on data.

What is "F" stand for?

Findable

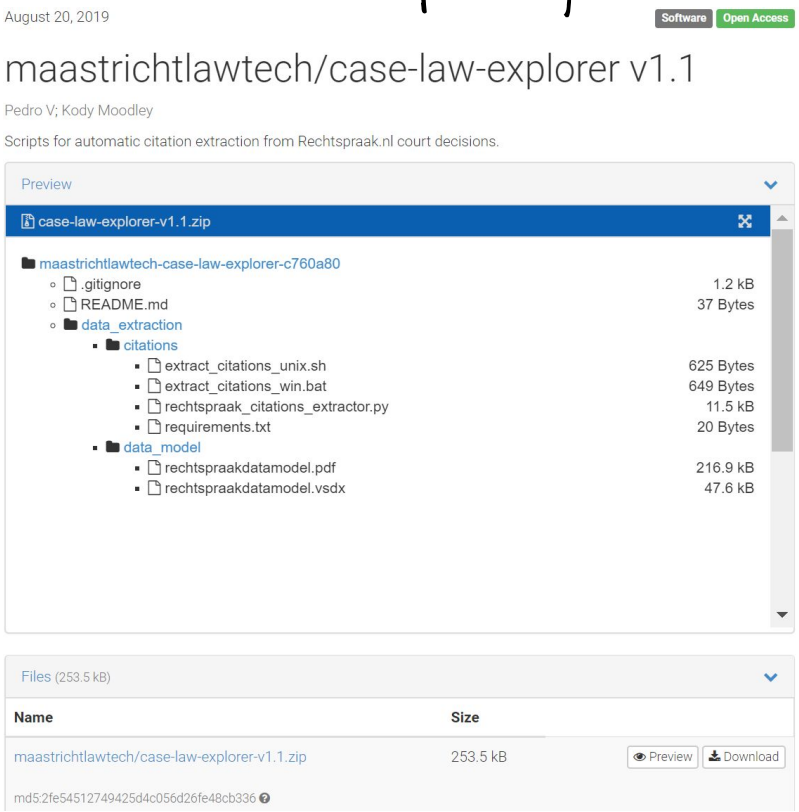
F1: (meta) data are assigned globally unique and persistent identifiers

Globally unique and persistent identifier for one resource in internet

Findable

https://doi.org/10.34894/FQGRKC

Git and Dataverse repository



Findable

Globally unique and persistent identifier for one particular person on the planet earth

Findable

https://orcid.org/0000-0003-4727-9435

iD ORCID
Connecting Research and Researchers

enables **unambiguous** identification of resources of interest.

Findable

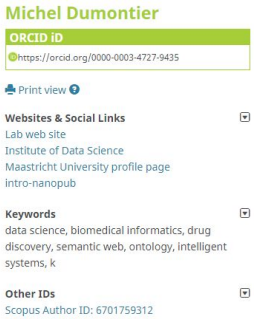
Namespace

https://orcid.org/0000-0003-4727-9435

scheme / protocol

domain name

resource name



Findable

F2: Data are described with rich metadata

Metadata and data

Maastricht University
Institute of Data Science

Findable

Data	Metadata
	**Collection:** Land North and South of Blundell's Road, Tiverton **Title:** General shot of trench 11, looking East **File name:** General_Shot_of_Trench_11_Looking_East **Created:** 2014-12-09 **Copyright:** Cotswold Archaeology **Creator:** Cotswold Archaeology **Coordinates:** OSGB 298080E 113000N **Location:** Blundell's Road; Tiverton; Devon; England **Subject:** Trench **Period:** Late Neolithic; Middle Bronze Age; Roman **Data Type:** Raster Image **File Type:** JPEG File Interchange Format **Software:** Canon AOS (version 6) **Checksum:** 38e27eba5c21ded56b51b8db8dbe0adc

Metadata and data

Maastricht University
Institute of Data Science

Data	Metadata
	**Collection:** Land North and South of Blundell's Road, Tiverton **Title:** General shot of trench 11, looking East **File name:** General_Shot_of_Trench_11_Looking_East **Created:** 2014-12-09 **Copyright:** Cotswold Archaeology **Creator:** Cotswold Archaeology **Coordinates:** OSGB 298080E 113000N **Location:** Blundell's Road; Tiverton; Devon; England **Subject:** Trench **Period:** Late Neolithic; Middle Bronze Age; Roman **Data Type:** Raster Image **File Type:** JPEG File Interchange Format **Software:** Canon AOS (version 6) **Checksum:** 38e27eba5c21ded56b51b8db8dbe0adc

Findable

Good metadata helps others to identify the your digital resources quickly via repositories for example

High quality with 300 Metadata fields? Any one got time for that?

Maastricht University
Institute of Data Science

Findable

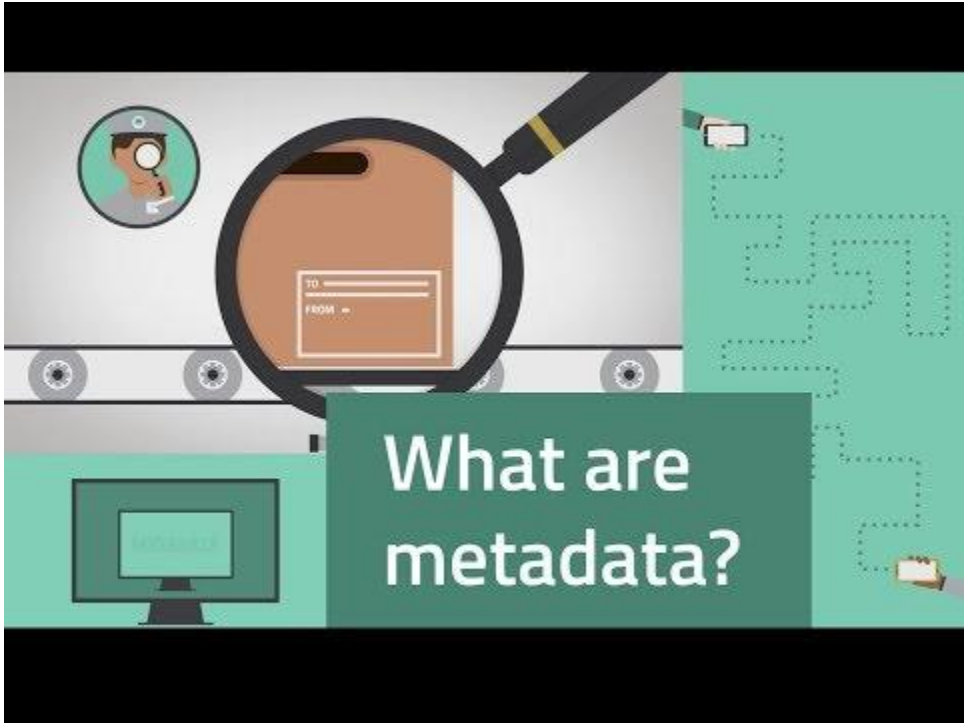
- What is the name or title of the digital resource?
- What is the digital resource about?
- Who contributed to creating or maintaining the digital resource?
- When was it created, modified, released?
- What methodology or tool was used?
- Which language is used?
- Which formats is it available as?
- What license is it released under?
- Which descriptive or quality metrics are available?
- Who is using it?

https://www.pinterest.com/pin/298645019032711046/

Why are metadata (and why are they so important?)

Maastricht University
Institute of Data Science

Findable



https://www.youtube.com/watch?v=-4_MFhi4GpU (2 min)

Barriers of Metadata

- Sounds like a ton of work - too time consuming
- We never learned how to share data
- Many initiatives seem to have very little buy-in from researchers
- Not very enticing to enter tons of metadata into a form in the vague hope that someday this will become easier to find for someone else
- We need some reward for the time and effort spent
- Increased utility of your data + methods

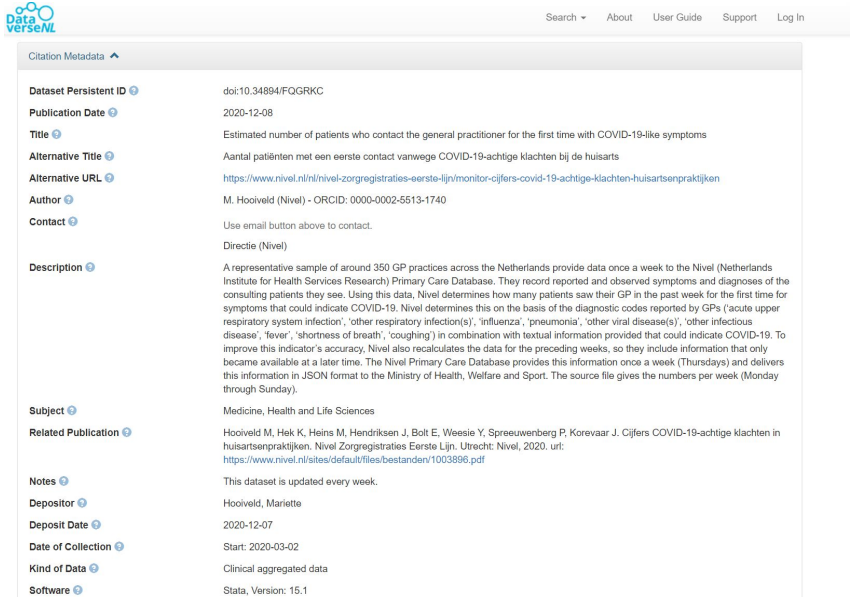
Preparing data is too time-consuming

Findable

F3: Metadata specify the data identifier

@c.utrillaguerrero@maastrichtuniversity.nl

Because good research need good data following best practices



DataCite

Mandatory Properties

Findable

Mandatory and Recommended properties and sub-properties are especially valua information seekers and added-service providers, such as indexers. The DataCite I Working Group members strongly urge the inclusion of metadata identified as Re the purpose of achieving greater exposure for the resource's metadata record, an underlying research itself.

Mandatory Properties
Identifier
Creator
Title
Publisher
PublicationYear
ResourceType

Recommended Properties

Recommended Properties
Subject
Contributor
Date
RelatedIdentifier
Description
GeoLocation

Optional Properties

Optional Properties
Language
AlternateIdentifier
Size
Format
Version
Rights
FundingReference

Findable

F4: (Meta)data are registered or indexed in a searchable resource

F3: Metadata clearly and explicitly include the identifier of the data they describe
(basically it states data and metadata should be in different files)

Maastricht University
Institute of Data Science

Supports manual exploration - twitter vaccination dataset example

Findable

Google

Dataset Search

Search for Data Sets

Try coronavirus covid-19 or education outcomes site:data.gov.

Learn more about Dataset Search.

Think: How does your discipline share data?

Findable

Illustration credit: Ainsley Seago.

https://www.nature.com/sdata/policies/repositories

Kept safe in a secure environment

Following a year-long public discussion and building on existing community consensus[1], several stakeholders, representing various segments of the digital repository community, have collaboratively developed and endorsed a set of guiding principles to demonstrate digital repository trustworthiness. **T**ransparency, **R**esponsibility, **U**ser focus, **S**ustainability and **T**echnology: the TRUST Principles provide a common framework to facilitate discussion and implementation of best practice in digital preservation by all stakeholders.

https://www.nature.com/articles/s41597-020-0486-7

What is "A" stand for?

Accessible 

Here the focus is on the 'ability' and 'conditions about compliance'

Accessible 

A1: (meta) data are retrievable by their identifier using a standard communications protocols.

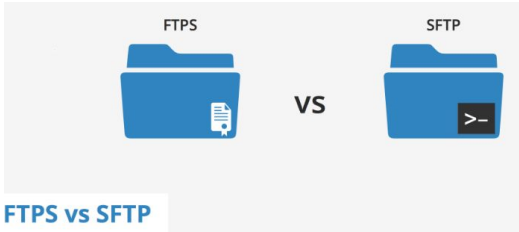
A1.1: the protocol is open, free, and universally implementable.

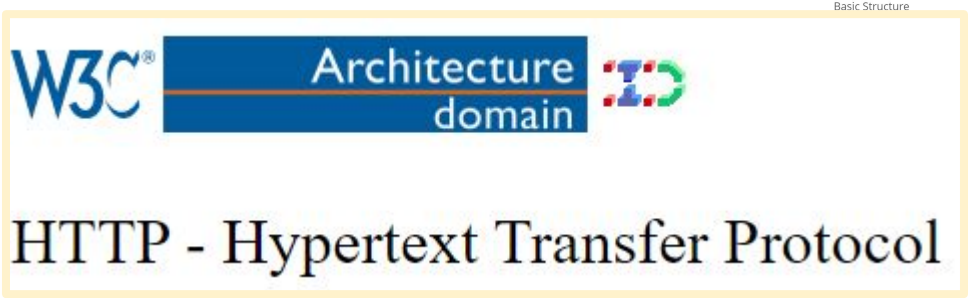
A1.2: the protocol allows for an authentication and authorisation procedure, where necessary.

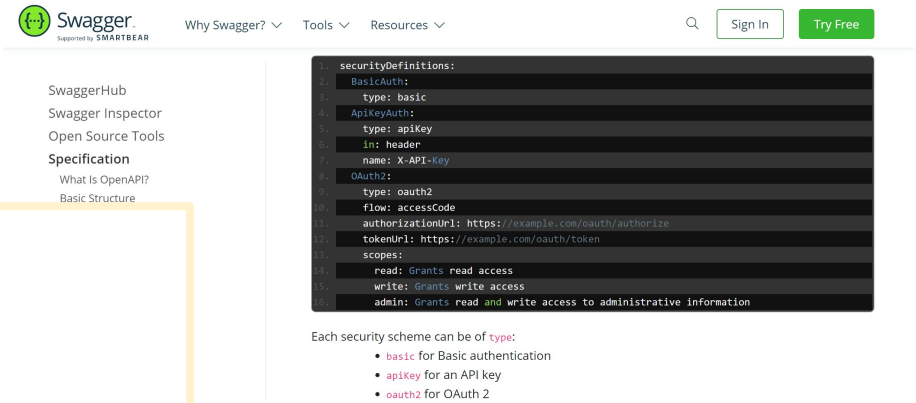
'Ability' = there should be no additional barriers to retrieve data

A.I.I. Several open, free and universally implementable protocols framework but HTTP protocol widely supported by community

Accessible



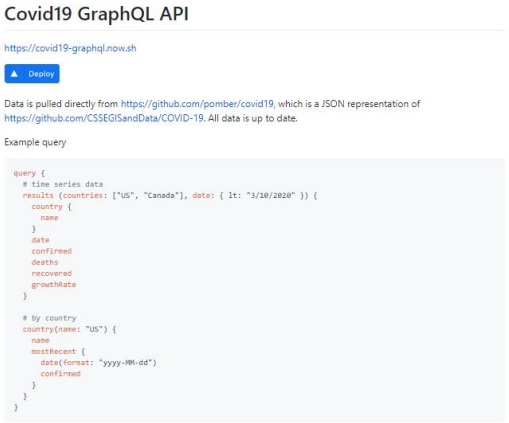




Why do we care about machine accessibility?

Accessible

- Teach computers how communicate and quickly access to data related to the pandemic - EU Open source solutions to help medical staff, public administrations, citizens in daily activities.



Why do we care about machine accessibility?

Accessible

ELSEVIER

What's an API? 5 things you need to know to stay current

APIs are having a growing impact on our lives and work – even if you're not a techie

By Alice Atkinson-Bonasio Posted on 24 September 2014

"Our users are researchers, and generally they spend a lot of time finding the right content, and then organising and formatting their data. We want to support our dev community so they use the Mendeley API to integrate our product not only with Elsevier's data and offerings, but those of other publishers – and any product out there that suits their needs and makes their lives easier."

@c.utrillaguerrero@maastrichtuniversity.nl

Why do we care about machine accessibility?

Accessible

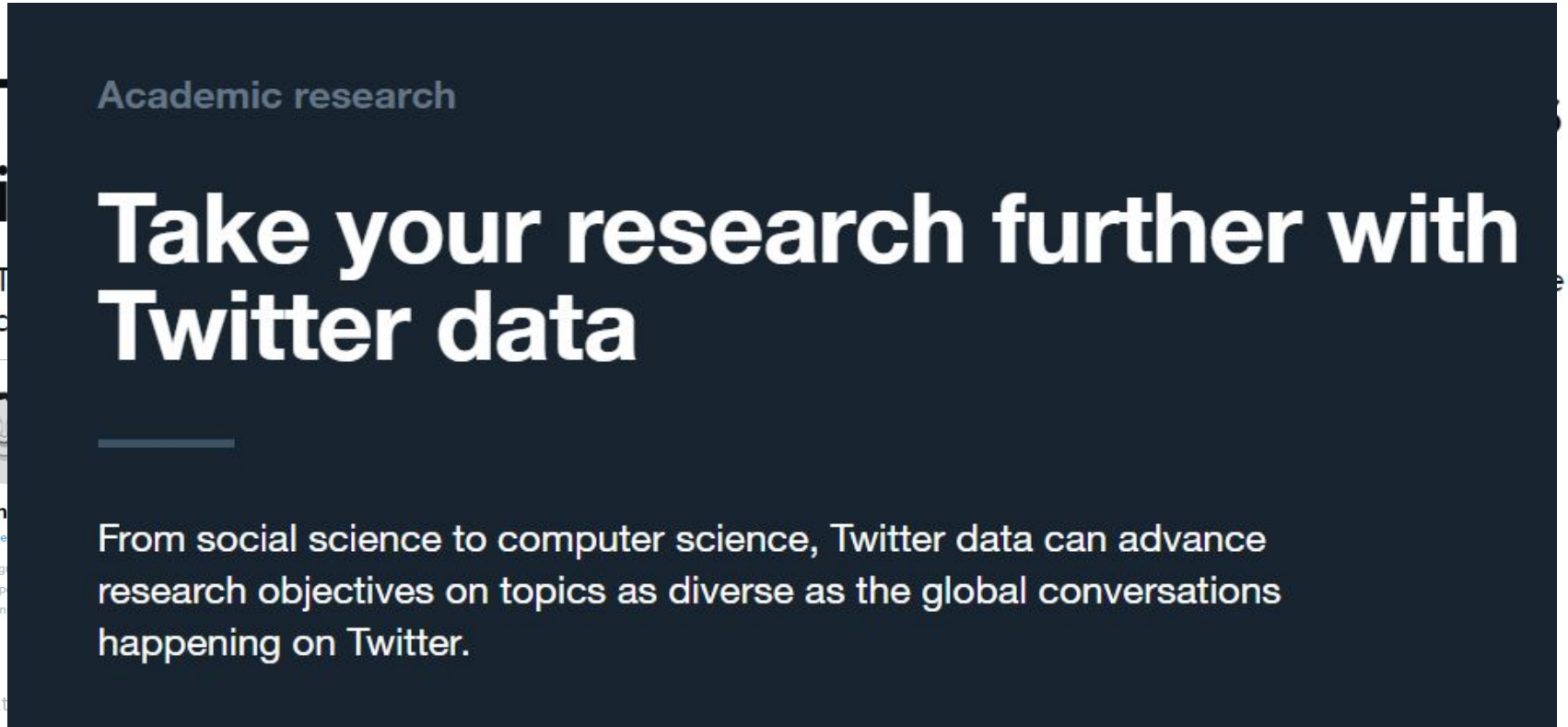
ELSEVIER

Interoperability is about ensuring that applications, tools and data sets from different providers can work together. The Mendeley API represents our commitment to interoperability with any tools that researchers need.

@c.utrillaguerrero@maastrichtuniversity.nl

Why do APIs matter to me?

Accessible



When to create an API?

- Collaboration and sharing your model with data scientist and software developers.
- Your data set is large, making download via FTP complicated.
- Your research community only need access to a part of the data at any one time.
- Your data changes or is updated frequently.
- Capture and share time-series data.
- Unstructured data to convert into structured format (e.g. *Linguistics*).
- Nice article for historian programming: why to use API as a researcher? And Europeana developer for cultural heritage.

@c.utrillaguerrero@maastrichtuniversity.nl

Why *open, free, universally implementable*?

Accessible

CORE TRUST SEAL

AI.2 for FAIR != "open"

Accessible

Most of the ESA Earth observation datasets are available on the Internet free of charge. Access is granted after user registration, which also provides the detailed content of the free datasets.

Sensitive Data Community of Practice

Accessible

- Government Data
- Health Data
- Indigenous Data
- Commercial Data

CARE Principles for Indigenous Data Governance

esa

Defense-security high protected data, but still FAIR

As open as possible, as closed as necessary

Accessible

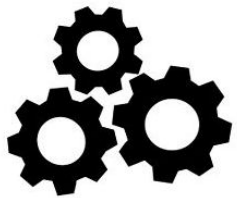
API OPTIONS		RF		
Request-driven API	RESTful HTTPS (JSON)	✓		
	RESTful HTTPS (XML)	✓		
Event-driven API	HTTPS Websocket (JSON)	✓	✓	✓
Export Formats	JSON, XML, SHP, KML	✓	✓	✓

esa

API = Application Programming Interface

What is "I" stand for?

Interoperable

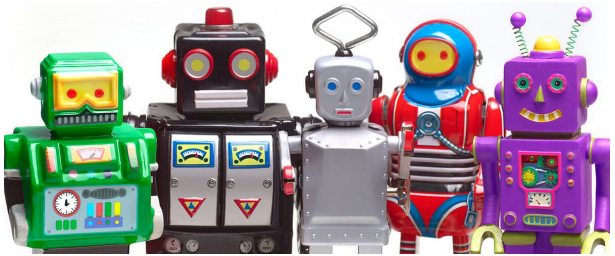


Interoperable 

I1: (Meta)data use formal, accessible, shared, and broadly applicable language for knowledge representation

Here we start empathizing with machines, and before!!

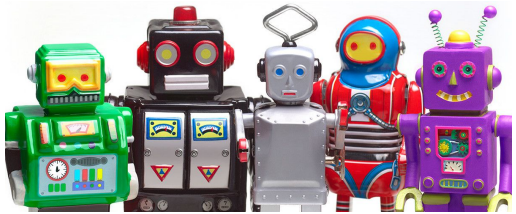
Be inclusive with others "common understanding"



inclusive = empower these agents by providing an environment that is understandable for them and easily make use of our data

Which is the primary goal of FAIR?

Step towards a common understanding of the data by machines, which is a prerequisite for a functional internet of FAIR Data - Interpretations 2020



Suppose a machine visits two datasets in which field "temperature" is present

Interoperable 

Date	Temp
01/01/2021	191
02/01/2021	195
03/01/2021	200
04/01/2021	199
05/01/2021	180

Machines and also humans need contextual information to determine things rather than assume

Interoperable

Date *(DD/MM/YY)*	Temp (*Kelvin*)
01/01/2021	191
02/01/2021	195
03/01/2021	200
04/01/2021	199
05/01/2021	180

```
{
 "coord": {
  "lon": -122.08,
  "lat": 37.39
 },
 "weather": [
  {
   "id": 800,
   "main": "Clear",
   "description": "clear sky",
   "icon": "01d"
  }
 ],
 "base": "stations",
 "main": {
  "temp": 282.55,
  "feels_like": 281.86,
  "temp_min": 280.37,
  "temp_max": 284.26,
  "pressure": 1023,
  "humidity": 100
 },
 "visibility": 16093,
 "wind": {
  "speed": 1.5,
  "deg": 350
 },
 "clouds": {
  "all": 1
 },
 "dt": 1560350645,
 "sys": {
  "type": 1,
  "id": 5122,
  "message": 0.0139,
  "country": "US",
  "sunrise": 1560343627,
  "sunset": 1560396563
 },
 "timezone": -25200,
 "id": 420006353,
 "name": "Mountain View",
 "cod": 200
}
```


GeoJSON

Even better for machine-readable i.e. value-key pair"

Interoperable

Fields in API response

- `coord`
  - `coord.lon` City geo location, longitude
  - `coord.lat` City geo location, latitude
- `weather` (more info Weather condition codes)
  - `weather.id` Weather condition id
  - `weather.main` Group of weather parameters (Rain, Snow, Extreme etc.)
  - `weather.description` Weather condition within the group. You can get the output in your language. Learn more
  - `weather.icon` Weather icon id
- `base` Internal parameter
- `main`
  - `main.temp` Temperature. Unit Default: Kelvin, Metric: Celsius, Imperial: Fahrenheit.

FAIR speaks to the ability of data to be reused by a generic agent, rather than a community-specific agent.

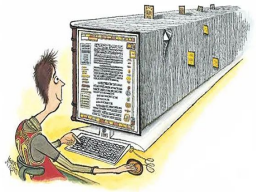
Maastricht University
Institute of Data Science

Interoperable

"The most widely-accepted choice is the Resource Description Framework (RDF) which is the W3C's recommendation for how to represent knowledge on the Web in a machine-accessible format" .
(FAIR principles Interpretation, 2021)

https://www.mitpressjournals.org/doi/full/10.1162/dint_r_00024

W3C Semantic Web



Interoperable 

I2: (Meta)data use vocabularies that follow the FAIR principles

Terminology you use, for instance, the units of measure, classifications, and relationship definitions are themselves **FAIR**

Interoperable

schema.org

FAIRsharing.org
standards, databases, policies

Insufficient information to enable a **machine** understand the meaning and context

Interoperable 

Date	Thing	Number
01/01/2020	Kiwi	1
02/01/2020	Kiwi	2
03/01/2020	Kiwi	3
04/01/2020	Kiwi	2
05/01/2020	Kiwi	1

Shared vocabularies

Interoperable

What do you mean by “Kiwi”?

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

https://en.wikipedia.org/wiki/Kiwi_(bird)

Chose the one most suitable for your dataset

Interoperable 

Date (DD/MM/YYYY)	Species	Number
01/01/2020	Kiwi	1
02/01/2020	Kiwi	2
03/01/2020	Kiwi	3
04/01/2020	Kiwi	2
05/01/2020	Kiwi	1

Interoperable 

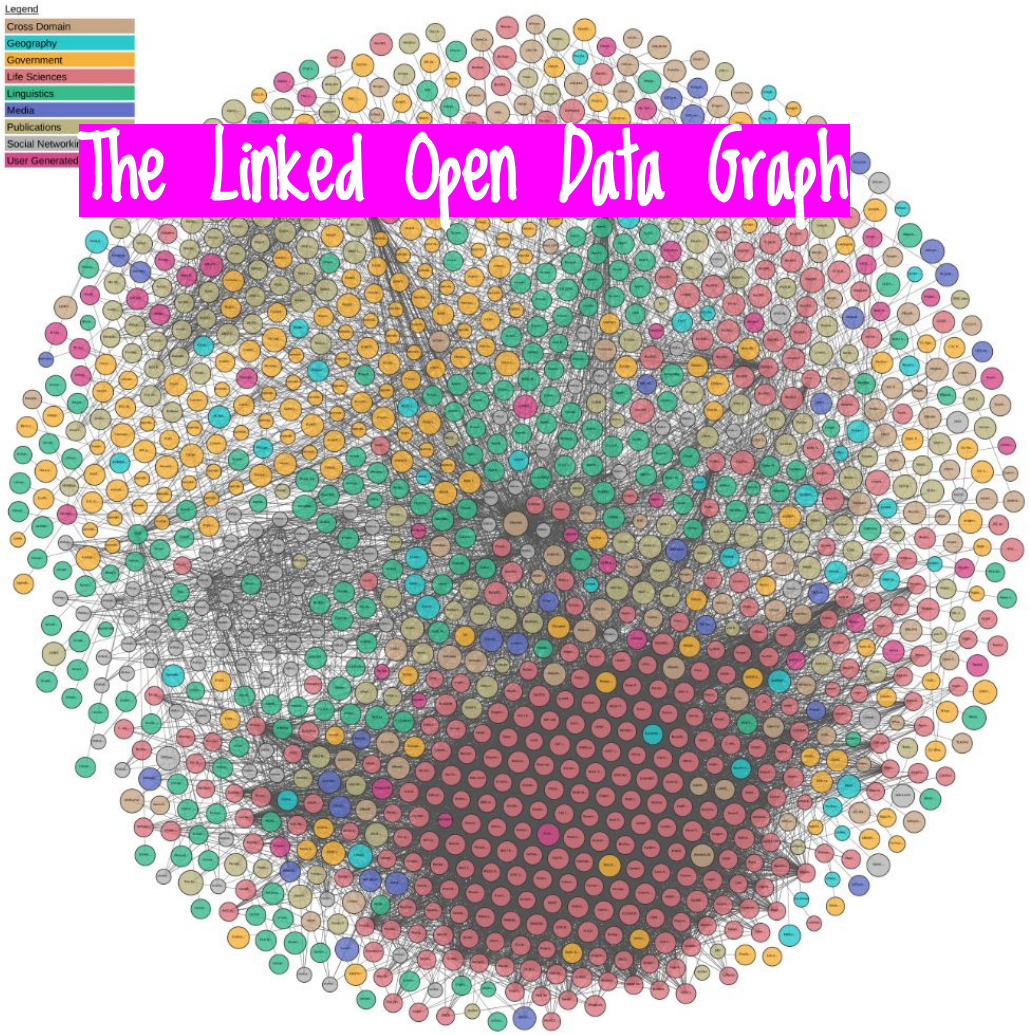
I3: (Meta)data include qualified references to other (meta)data

Maastricht University
Institute of Data Science

Interoperable

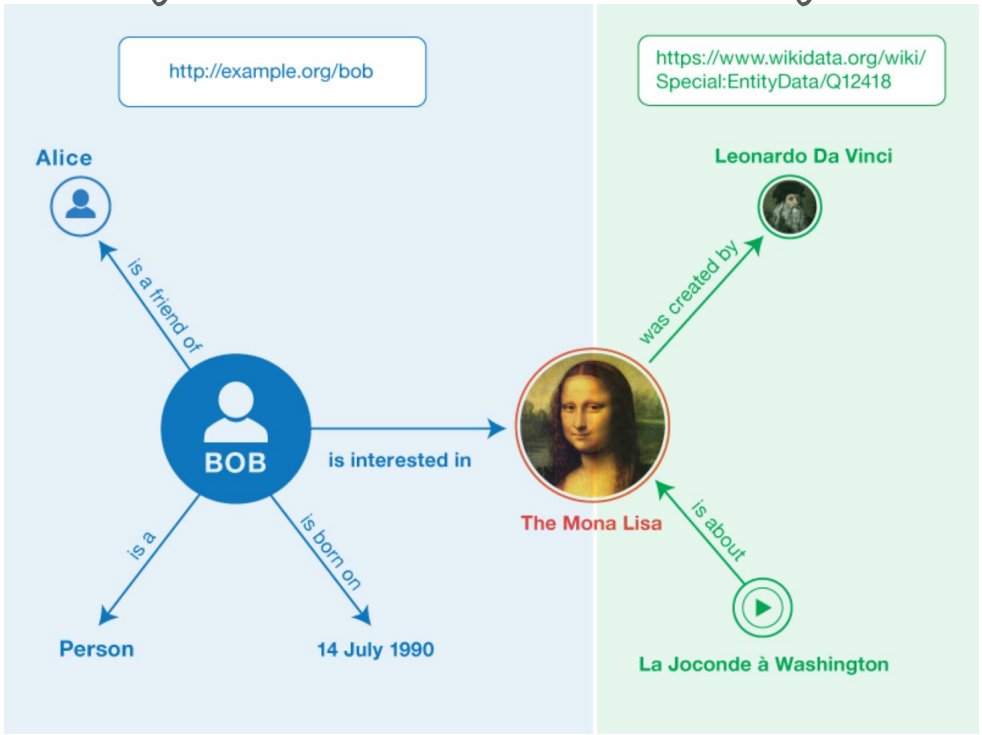
Interoperable

The Linked Open Data Graph



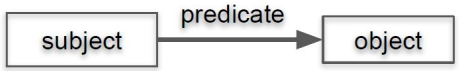
1,224 datasets
with 16,113 links
(as of June 2018)

Subjects, Predicates and Objects



Interoperable

RDF Data Model: A set of **triples** that consists in **three** components:



The triple expresses a relationship between the subject and the object. The subject and object represent nodes in the RDF graph
and the predicate represents an edge / relation between these nodes

Knowledge Graph

Interoperable

What is "R" stand for?

R1.1. (Meta)data are richly described with a plurality of accurate and relevant attributes

wait..It's not similar to Findable Principle (F2)?

Reusable ♻

We chose the term *'plurality'* to indicate that the metadata author should be as generous as possible in providing metadata, even including information that may seem irrelevant. - GO FAIR

F2: Repository and indexed metadata.
R1: Is your data valid for my study? Give me instructions (i.e. readme) to assess if your data is appropriate for me.

R1.1. (Meta)data are related with clear and accessible data usage license

Reusable

Licenses & Standards

About Open Source Licenses

Open source licenses are licenses that comply with the Open Source Definition — in brief, they allow software to be freely used, modified, and shared. To be approved by the Open Source Initiative (also known as the OSI), a license must go through the Open Source Initiative's license review process.

open source initiative
Approved License®

Popular Licenses

The following OSI-approved licenses are popular, widely used, or have strong communities:

- Apache License 2.0
- BSD 3-Clause "New" or "Revised" license
- BSD 2-Clause "Simplified" or "FreeBSD" license
- GNU General Public License (GPL)
- GNU Library or "Lesser" General Public License (LGPL)
- MIT license
- Mozilla Public License 2.0
- Common Development and Distribution License
- Eclipse Public License version 2.0

https://opensource.org/licenses

creative commons

Choose an open source license

An open source license protects contributors and users. Businesses and savvy developers won't touch a project without this protection.

Which of the following best describes your situation?

I need to work in a community.

Use the license preferred by the community you're contributing to or depending on. Your project will fit right in.

If you have a dependency that doesn't have a license, ask its maintainers to add a license.

I want it simple and permissive.

The MIT License is short and to the point. It lets people do almost anything they want with your project, like making and distributing closed source versions.

Babel, .NET Core, and Rails use the MIT License.

I care about sharing improvements.

The GNU GPLv3 also lets people do almost anything they want with your project, except distributing closed source versions.

Ansible, Bash, and GIMP use the GNU GPLv3.

R1.2. (Meta)data are associated with detailed provenance

Reusable

A) Organized by File type

Example.A
- Code
  - Step.1
  - Step.2
- Data
  - Processed
  - Raw
- Results
  - Figure.1
  - Figure.2
  - Models
- readme.txt

B) Organized by Analysis

Example.B
- Figure.1
  - Code
  - Data
  - Results
- Figure.2
  - Code
  - Data
  - Results
- readme.txt

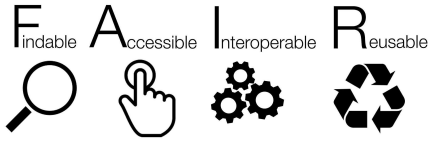
Facilitate reproducibility by reproducible research things with community standards

There's a README.md as part of the git repository that can give you some background. A Institute generic version of this can be found in a branch on the github.

Lessons

Lesson 1 - Documentation
Lesson 2 - Naming Conventions
Lesson 3 - Folder Structure
Lesson 4 - Automation
Lesson 5 - Version control
Lesson 6 - Cloud Backups
Lesson 7 - Computer Security
Lesson 8 - Separating Identified Variables
Lesson 9 - Permanent Identifiers for your Published Results

Summary FAIR principles



Findable: Globally unique, persistent identifier (e.g., DOI), rich metadata, indexed for search

Accessible: Retrievable using standardised, open protocol (e.g., HTTPS). Metadata stay accessible when data is removed.

Interoperable: Metadata use a formal, accessible, shared, broadly applicable language/vocabulary,references to other metadata

Re-usable: Accurate and relevant attributes, provenance and data usage licence is clear, meet domain-relevant community standards

Why go FAIR? [write 3 pros]

- To be able to use your data it should be **findable** and **accessible**
- Make it easier for me to **use my data for a new purpose**
- For effective **reuse** since all relevant aspects are reported and it has a detailed license and provenance
- Make it easier/possible for people to **verify my work**
- Ensure the data are available in the future, especially after I stopped caring about them
- Making data interoperable allows for better analysis
- Satisfy expectations around data management from institutions

Why practice FAIR objects? [write 3 pros]

The idealist

- **Shoulders of giants!**
- Validates scientific knowledge
- Allows others to build on your findings
- Improved transparency
- Increased transfer of knowledge
- Increased utility of your data + methods

The pragmatist

- Increased efficiency
- Reduces false leads based on irreproducible findings
- Data sharing citation advantage (Piwowar 2013)
- "It takes some effort to organize your research to be reproducible… the **principal beneficiary is generally the author herself**."- Schwab & Claerbout

Poll

Go to slido.com

Enter code: **FAIR-UM**

FAIR for your dataset

1. **Build and collect your dataset using existing standards:**
    a. Assign a meaningful variable name to every column
    b. Format the data using your community standards (e.g. csv, json, rdf)
    c. Capture the provenance and context of your dataset
2. **Create high quality metadata to document your dataset:**
    a. Use vocabularies and standards
    b. Use specific metadata terms where necessary
3. **Make your dataset available to others**
    a. Version your dataset, and publish the data and metadata e.g. DataverseNL
    b. Create access points via APIs if necessary