

# BioData.pt

## Ready for BioData Management?



# Introduction to Data Management in Science

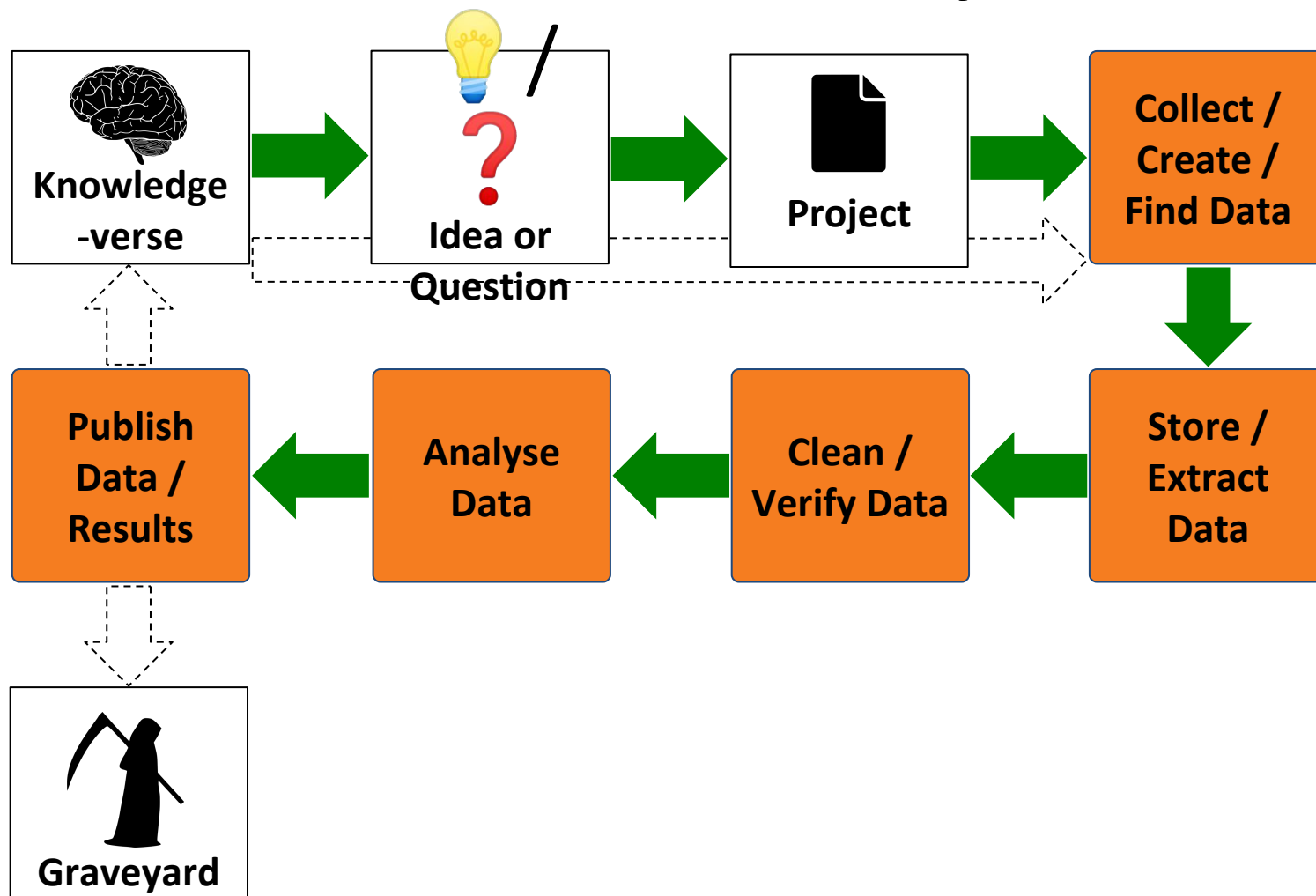
Daniel Faria



# 1. What Is Data Management?

# What Is Data Management?

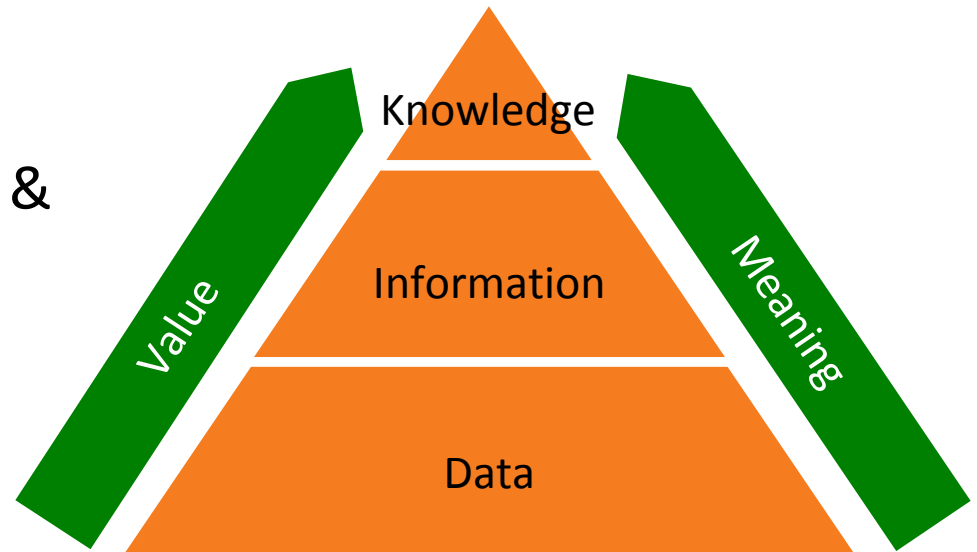
## Domain: The Data Lifecycle



# What Is Data Management?

## Goals:

- Facilitate data & knowledge discovery
- Avoid loss of data
- Ensure data accessibility & security
- Improve data processes
- Increase data value
- Allow data reuse



# What Is Data Management?

## Topics:

- Data Governance
- Data Architecture
- Data Modeling
- Data Storage & Maintenance
- Data Security
- Data Integration & Interoperability
- Documents & Content
- Data Analysis & Mining
- Metadata
- Data Quality

# Who Can Benefit From Data Management?

Anyone whose research requires data:

- analysis
- integration
- modelling
- processing
- production
- publication
- (re)use
- storage



By OnePoint Services, CC BY 2.0

<https://www.flickr.com/photos/78830297@N05/14740342074/>

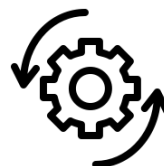
# What Are The Benefits Of Data Management?

## Improve research:



By LAFS,  
CC BY 2.0

**Effectiveness** –  
obtain more/better  
results



By Youmena,  
CC BY 2.0

**Efficiency** – improve  
productivity and  
cost-efficiency



By ROZMOWA,  
CC BY 2.0

**Security** – reduce  
data loss / control  
access to data



By Nithinan Tatah,  
CC BY 2.0

**Impact** – facilitate  
dissemination and  
knowledge discovery

All images: <https://thenounproject.com/>



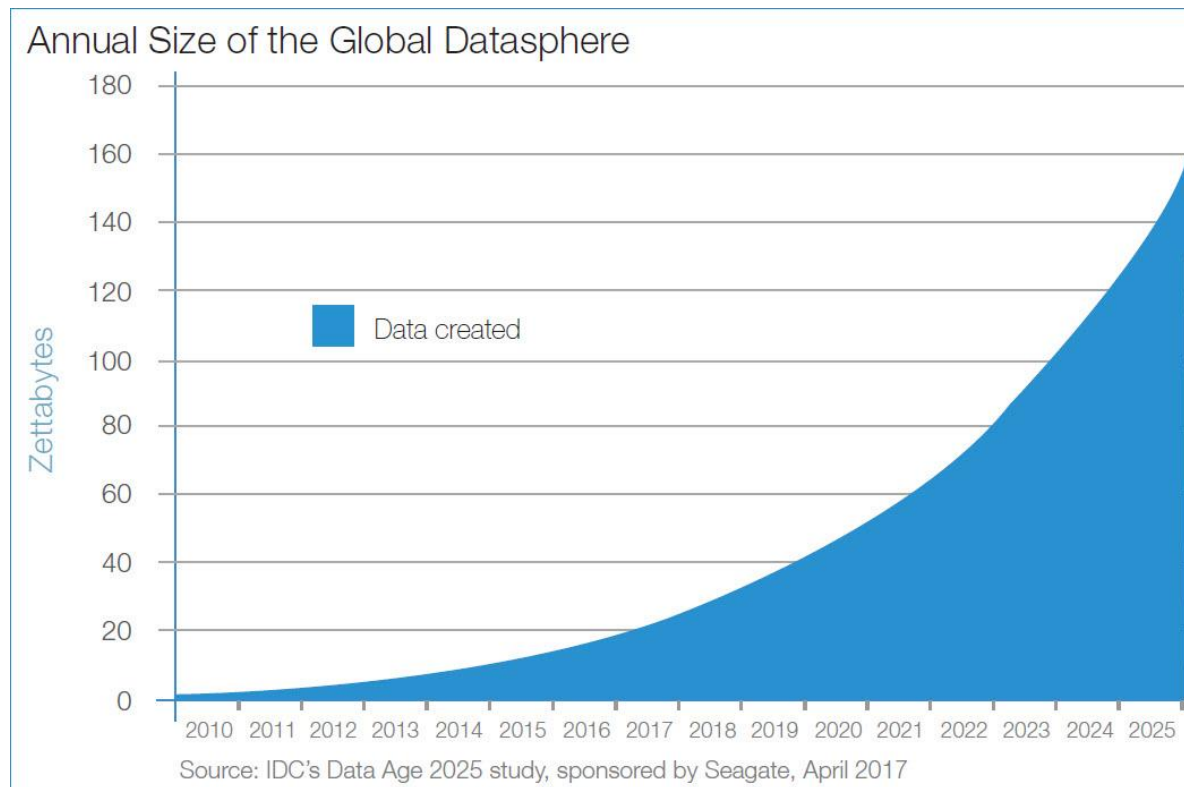
## 2. The Knowledge Discovery Problem



# The Knowledge Discovery Problem

## Exponential Growth Of:

- **Scientific Data**
  - Sequences
  - Structures
  - Tables
  - Images
- **Scientific Knowledge**
  - Software
  - Scientific Papers
  - Patents



# The Knowledge Discovery Problem

## Findability:

- More data  $\Rightarrow$  harder search
- Things can get lost amid a sea of things
- Efficient search is paramount
- If it is not findable, it may as well not exist

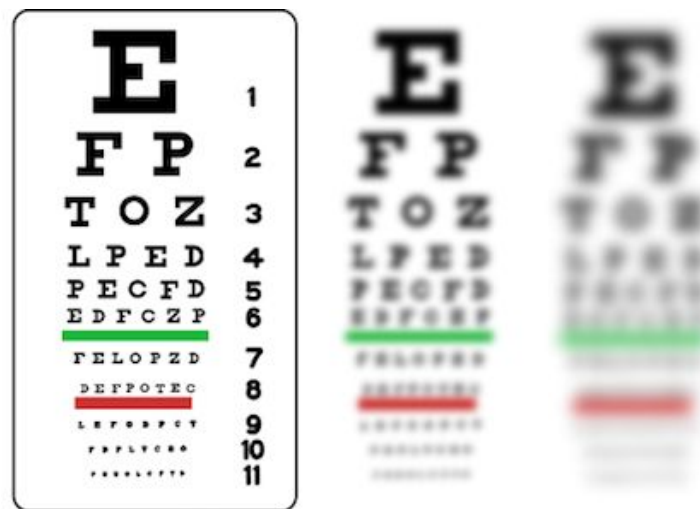


By Martin Handford, retrieved from:  
<https://exploringyourmind.com/how-does-our-brain-find-waldo/>

# The Knowledge Discovery Problem

## Interpretability:

- More data  $\Rightarrow$  less time to interpret
- We become myopic by necessity—can't afford the time to read the fine-print (e.g. a full research paper)
- If we cannot interpret it readily, then it is nearly useless

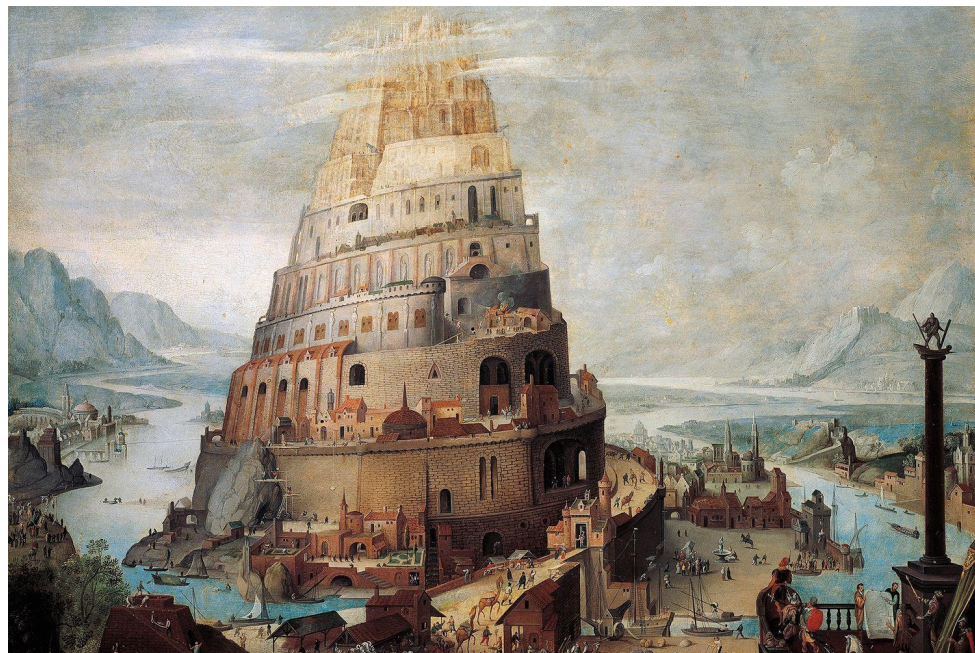


By Daniel P. B. Smith, CC BY-SA 3.0  
<https://en.wikipedia.org/wiki/File:Snellen-myopia.png>

# The Knowledge Discovery Problem

## Interoperability:

- More specialization  $\Rightarrow$  vocabulary and viewpoint divergence
- Reuse  $\Rightarrow$  interoperability  $\Rightarrow$  standardization
- If we use local dialects, our data and knowledge will be sundered



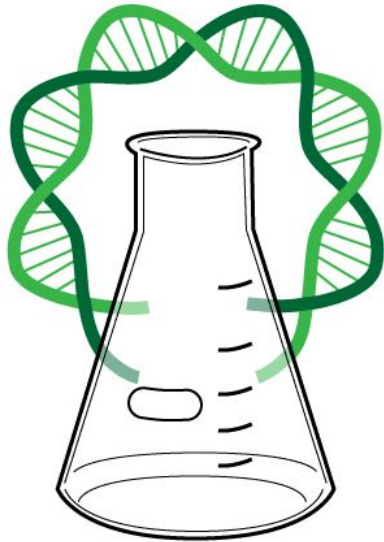
By Abel Grimmer, retrieved from:  
<http://cbcpnews.net/cbcpnews/the-tower-of-babel/>

# The Knowledge Discovery Problem

## Conclusions:

- Publishing research outputs only in scientific articles is not enough
  - Articles are not efficient vehicles for knowledge sharing!!!
- If we want our research outputs to effectively contribute to the knowledge-verse:
  - We must publish the data we produce
  - In a form that is easy to read
  - With sufficient information (metadata) to enable interpretation
  - In a form that is also easy to read





open science

By Greg Emmerich, CC BY-SA 3.0

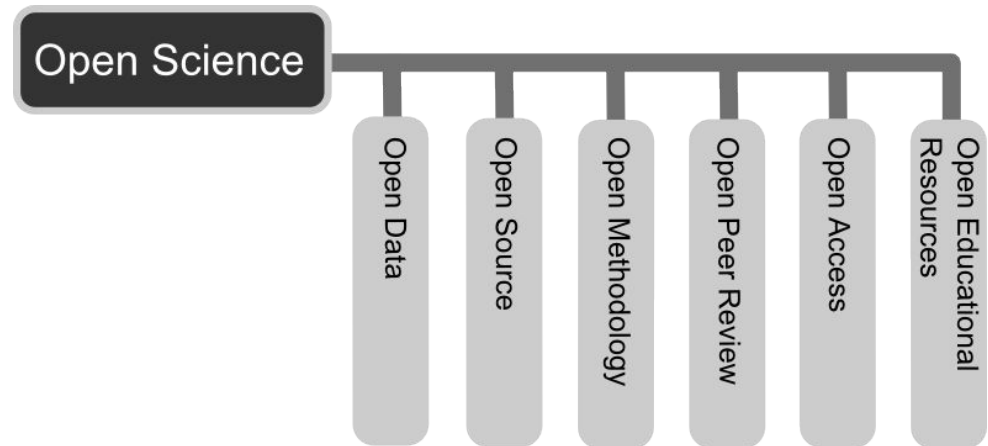
<https://www.flickr.com/photos/gemmerich/8732559382/>

## 3. What Is Open Science?

# What is Open Science?

## Goals:

- Scientific research and its dissemination accessible to all levels of society
  - publications
  - data
  - physical samples
  - software
  - ...
- Transparent and accessible knowledge shared and developed through collaborative networks



By Andreas E. Neuhold, CC BY 3.0

[https://en.wikipedia.org/wiki/File:Open\\_Science\\_-\\_Prinzipien.png](https://en.wikipedia.org/wiki/File:Open_Science_-_Prinzipien.png)

# What is Open Science?

## Topics:

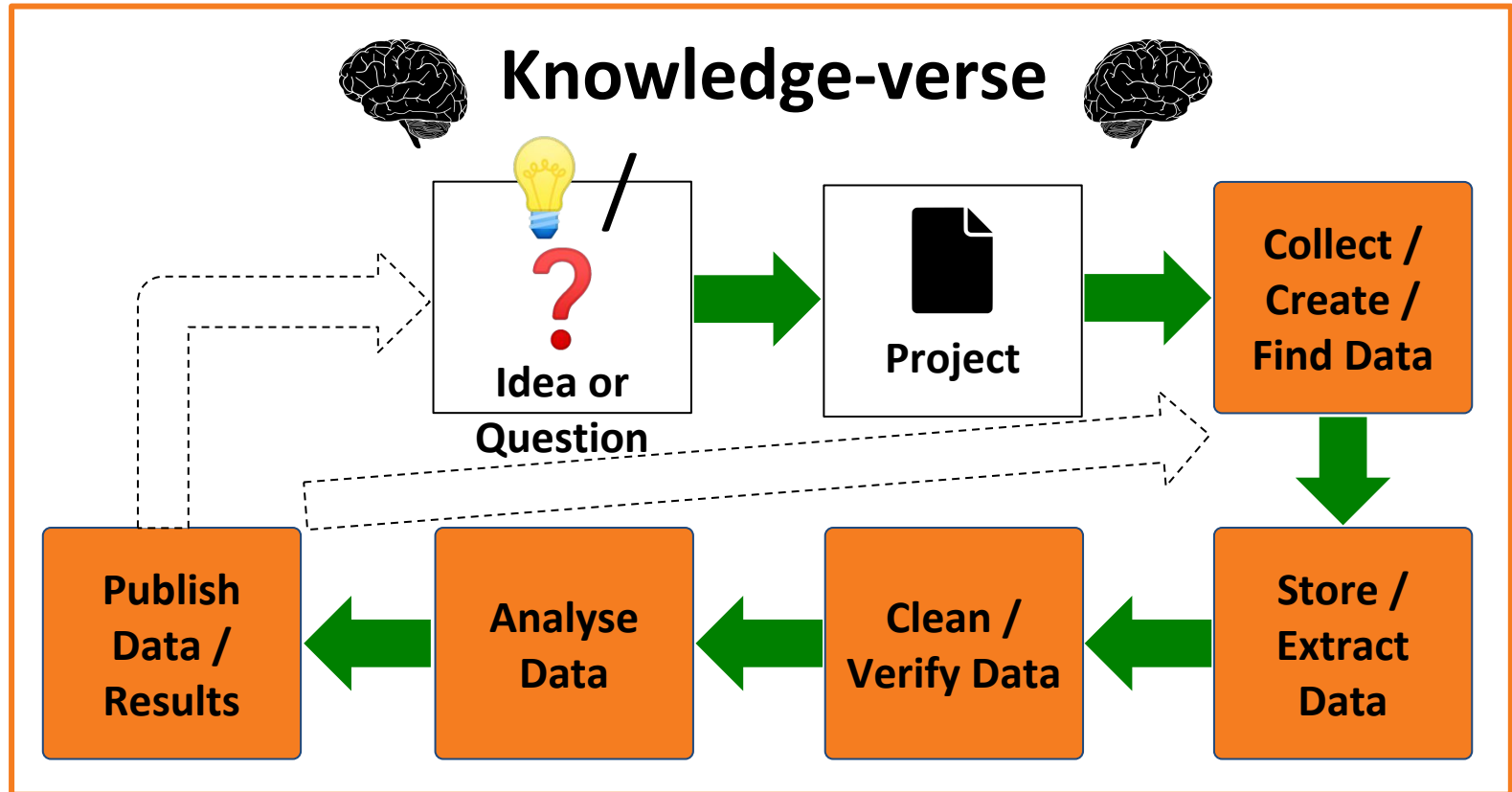
- **Open Access:** research outputs distributed online, free of cost or access barriers
- **Open Research:** data, result and methodology clearly documented and freely available online
- **Open-Notebook Science:** primary record of a research project publicly available online as it is recorded—no insider information



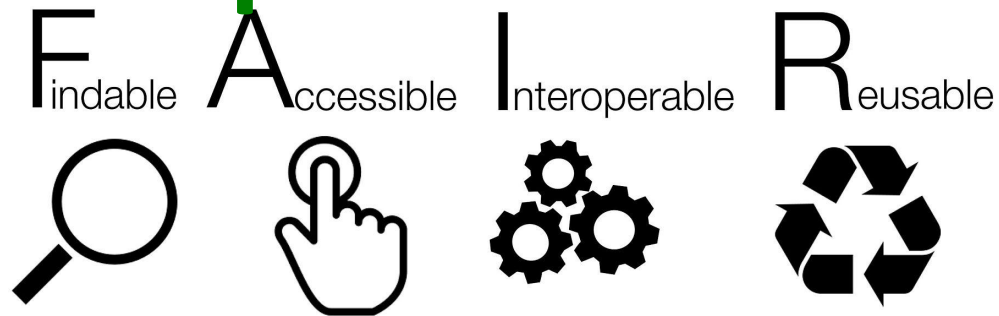


# What Is Open Science?

Everything documented & freely available!



# 4. What Are The FAIR Data Principles?



By Sangya Pundir, CC BY-SA 4.0

<https://commons.wikimedia.org/w/index.php?curid=53414062>

# What Are The FAIR Data Principles?

## Overview:

A set of four principles detailed in fifteen guidelines, that scientific data should aim to comply with.

**Findability** – (Meta)data should be easy to find for both humans and computers

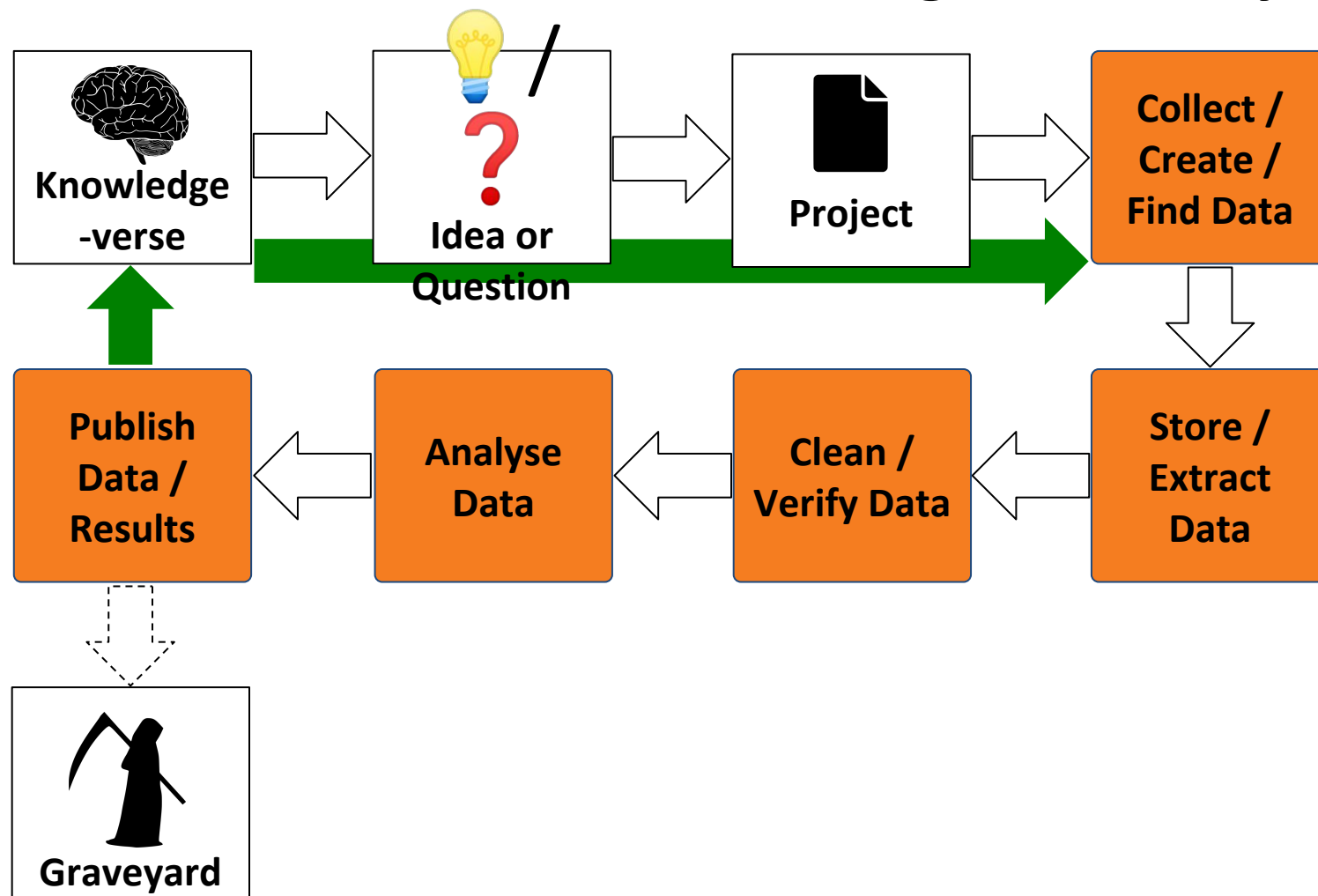
**Accessibility** – (Meta)data should have a defined access protocol with authentication and authorization rules

**Interoperability** – (Meta)data should be integratable with other similar datasets and interpretable by applications or workflows for analysis, storage, and processing

**Reusability** – (Meta)data should be well described so that it can be interpreted and reused

# What Are The FAIR Data Principles?

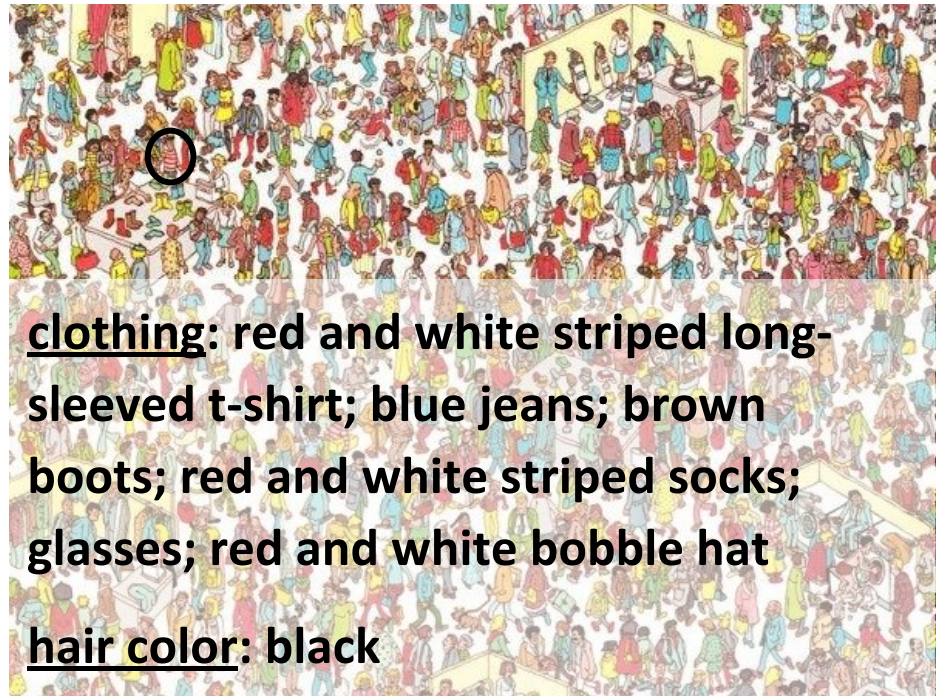
**Goal: Enable Data & Knowledge Discovery**



# Going FAIR Enables Knowledge Discovery

## Findability:

- Describe data with precise metadata useful for searching
- Use a common (structured) controlled vocabulary for metadata fields and values
- Put data in a repository that:
  - Uses persistent unique identifiers
  - Indexes metadata and allows searches



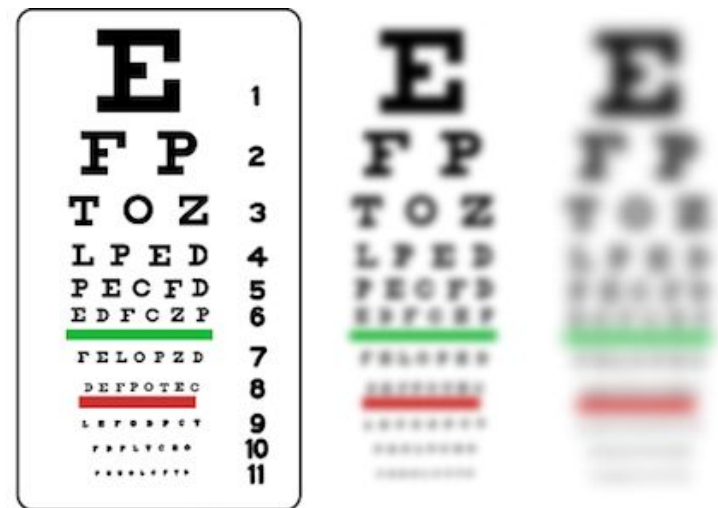
**clothing: red and white striped long-sleeved t-shirt; blue jeans; brown boots; red and white striped socks; glasses; red and white bobble hat**  
**hair color: black**

By Martin Handford, retrieved from:  
<https://exploringyourmind.com/how-does-our-brain-find-waldo/>

# Going FAIR Enables Knowledge Discovery

## Interpretability:

- Describe data with sufficient metadata for interpreting it and understanding the experimental context—each dataset should be fully self-contained
- Use a common (structured) controlled vocabulary for metadata fields and values



By Daniel P. B. Smith, CC BY-SA 3.0  
<https://en.wikipedia.org/wiki/File:Snellen-myopia.png>



# Going FAIR Enables Knowledge Discovery

## Interoperability:

- Use a common (structured) controlled vocabulary for metadata fields and values
- Include cross-references to external data objects whenever suitable (e.g. a sample in *BioSamples*)



By Abel Grimmer, retrieved from:  
<http://cbcpnews.net/cbcpnews/the-tower-of-babel/>

# Going FAIR Enables Knowledge Discovery

## Wait, That's FII, Not FAIR...

Indeed, but:

- Reusability is the end-goal, not the core problem—it hinges entirely on Interpretability and Interoperability.
- Accessibility is a technical problem—it concerns data repositories more than researchers—and it is already well addressed. As long as you publish your data in a well-established repository and define an authorization policy (when applicable, such as for sensitive data) you are well off.



# 5. To Be or Not to Be Open & FAIR???



[https://shakespeareoxfordfellowship.org/wp-content/uploads/Hamletskull\\_featured.jpg](https://shakespeareoxfordfellowship.org/wp-content/uploads/Hamletskull_featured.jpg)

# Pros & Cons

## Pros:









- Facilitates knowledge discovery
- Promotes reproducibility
- Enables networking
- Helps demystify science for the general public

## Cons:

- Care with sensitive data and with knowledge that has dangerous misuse potential
- Harder to make money off of your research
- Harder to stay ahead of your rivals
- Harder to fake it

# Pros & Cons

## FAIR & Open Data Payoff Matrix:

|                    | Me   | Others  |
|--------------------|--|---|
| Sitting on my data |    |    |
| Open data          |    |    |
| FAIR data          |    |  |

# FAQ

- **Can I receive credit for publishing data?**
  - You can. We are amidst a shift towards crediting data publishers as much as paper publishers.
- **Can't someone publish a paper ahead of me if I release my data?**
  - If someone can write a paper using your data ahead of you that supersedes yours, shame on you. If it does happen, you at least get credit for the use of your data, and will likely still be allowed to publish your paper as the original author of the data.
- **What if someone uses my data without giving me credit?**
  - The same can happen with paper publication. Reviewers and editors are expected to police this. Authors that do so can be red flagged.

# To Be or Not to Be Open & FAIR???

## It Helps Science!

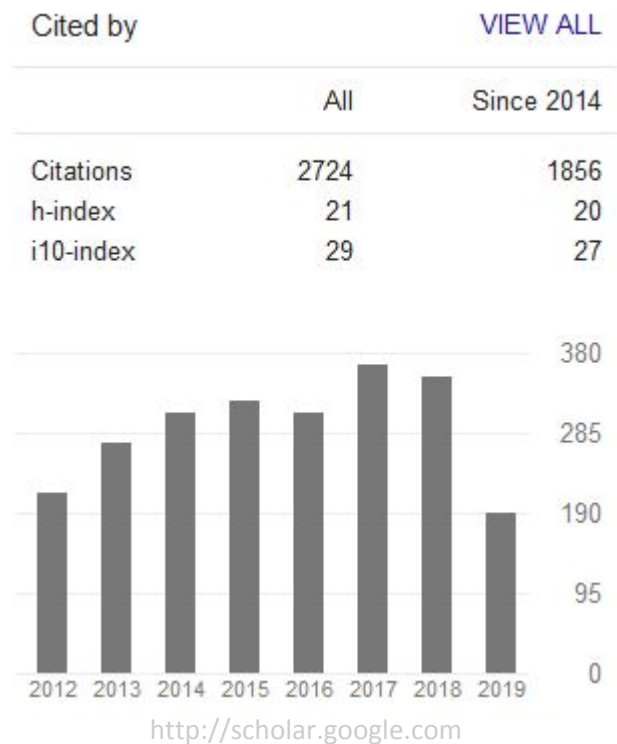
- Enables others to apply your knowledge in contexts beyond your foresight
- Enables others to reuse your data to make new research



# To Be or Not to Be Open & FAIR???

## It Helps You!

- It is easier to find and reuse your own data
- It is easier to write and submit a research paper
- If others apply or reuse your research, you get more citations (citing or crediting datasets is becoming common practice)



# To Be or Not to Be Open & FAIR???

## You'll Need It To Get Funded!

- Soon it will be impossible to get public funding in Europe without adherence to Open Science and FAIR
- FAIR compliance is starting to be verified
- A good track record will contribute to project approval



# To Be or Not to Be Open & FAIR???

## Only You Can Do It!

- Your research won't become FAIR unless you do it
- No one can make it FAIR for you
- Only you have the knowledge to describe your experimental setting, and organize and describe your data







## 6. FAIR Data Management

# FAIR Data Management

## It requires work!

- Preparing a dataset in accordance with the FAIR principles takes a lot of work:
  - Describe the experimental setting: context, goals, materials, data inputs
  - Describe the experimental procedure: protocols, data analysis procedures
  - Describe the data outputs
  - Annotate your (meta)data with ontologies

# FAIR Data Management

**It is hard if you do it only when publishing!**

- Have to trace all the data
  - Risk of data loss
- Have to recall/locate all the details about the experimental setting and procedures
  - Risk of metadata loss
- It is a lot of boring work to do at once
  - Inertia and rush lead to poor documentation

# FAIR Data Management

**It is easier if you do it from the start!**

- Make a Data Management Plan for your research!!!
- Document everything...
  - every breath you take, every move you make
- ...in a digital platform
  - Electronic lab notebook
  - Local shared repository with version control and clear structure (dropbox, google drive, shared local hard drive with git)

# FAIR Data Management

**It is easier if you do it from the start!**

- Publish everything that can be published mid-experiment
  - Software code, analysis pipelines, protocols, even preliminary or intermediate data
  - <https://github.com/>, <https://fairdomhub.org/>,  
<https://zenodo.org/>, <https://dataverse.org/>
- Plan in advance for how you'll need to document your experiment
  - Get acquainted with applicable metadata standards, domain ontologies, data publication formats

# FAIR Data Management

## Some Steps Are Still Hard!

- The Biomedical Ontology landscape is complex and hard to navigate:
  - There are often overlapping ontologies for a given domain
  - And worse, the same concepts appear in several ontologies, sometimes with the same URI!!!
  - But there are also domains with no (suitable) ontology
- Metadata standards exist only for a few domains, and not all specify a data format for publication
- Generic data repositories (e.g. FAIRDOMHub, Zenodo, Dataverse) have rigid data models that are not compatible with all domains / standards

# FAIR Data Management

## Do the Best You Can!

- FAIRness is a spectrum, and halfway there is a step forward

## Reach Out For Help!

- Data stewards and data management experts can provide guidance

## Things Will Get Easier!

- There are people working towards more user-friendly data management solutions—they need feedback on what can be improved