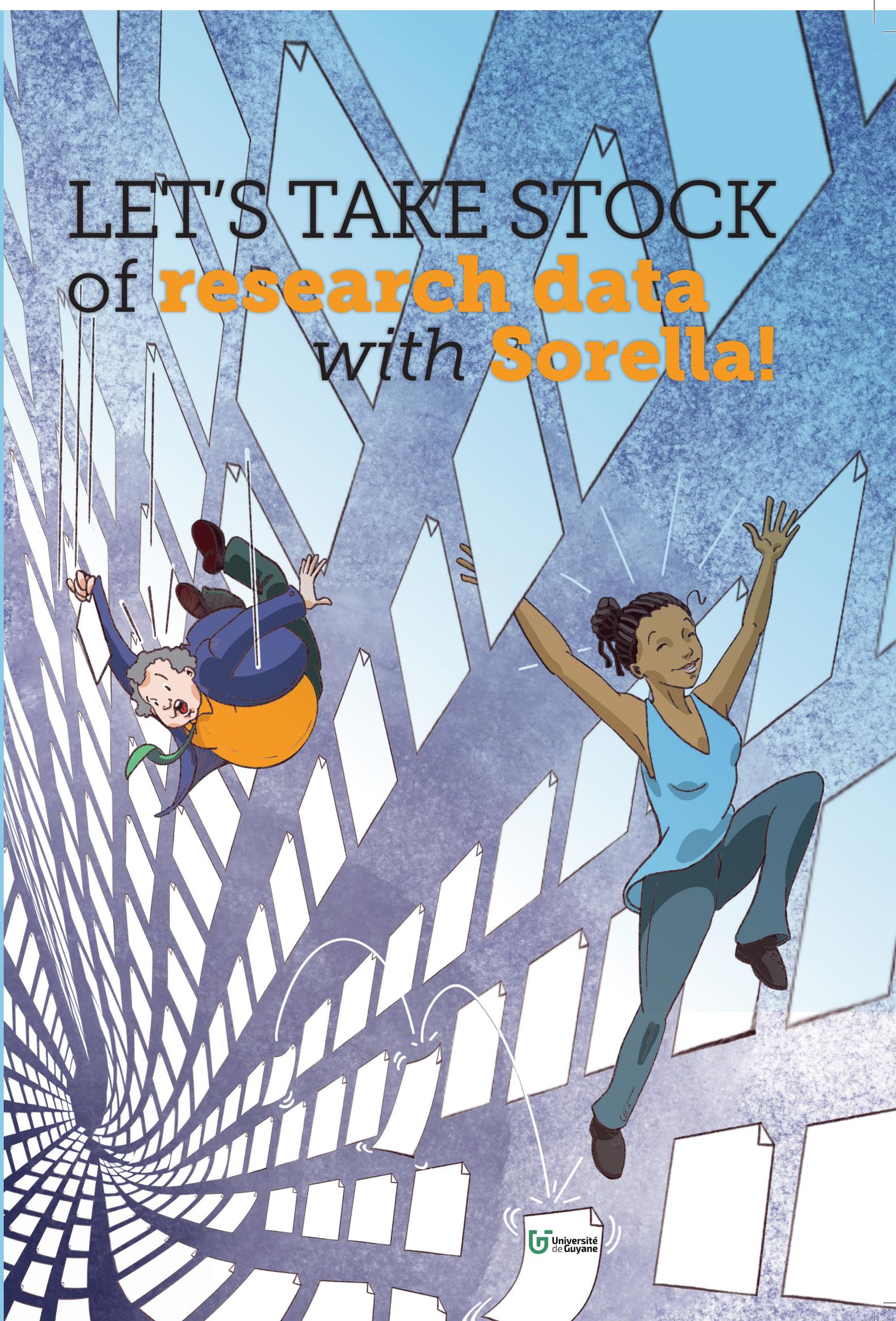
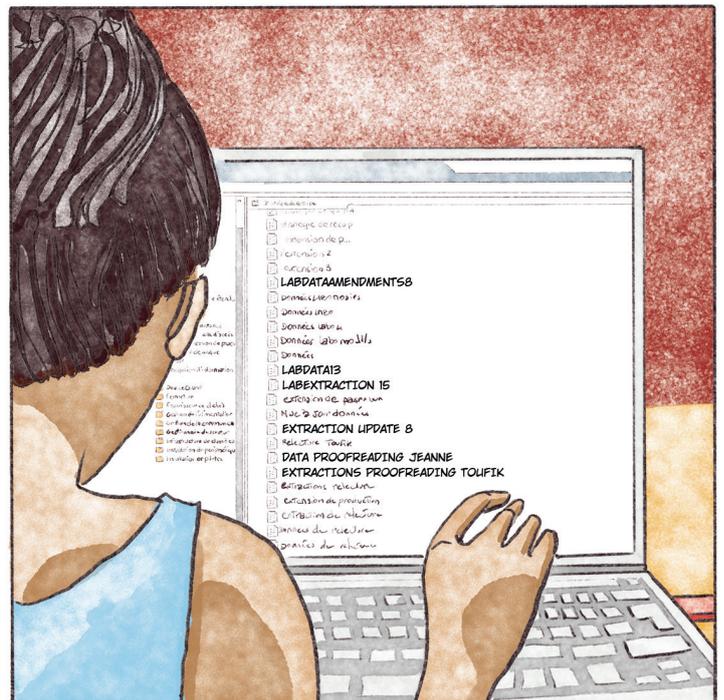
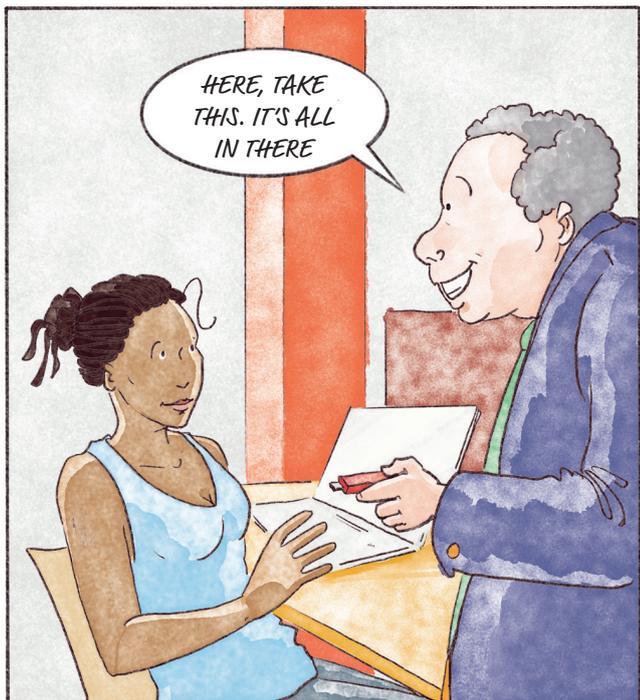
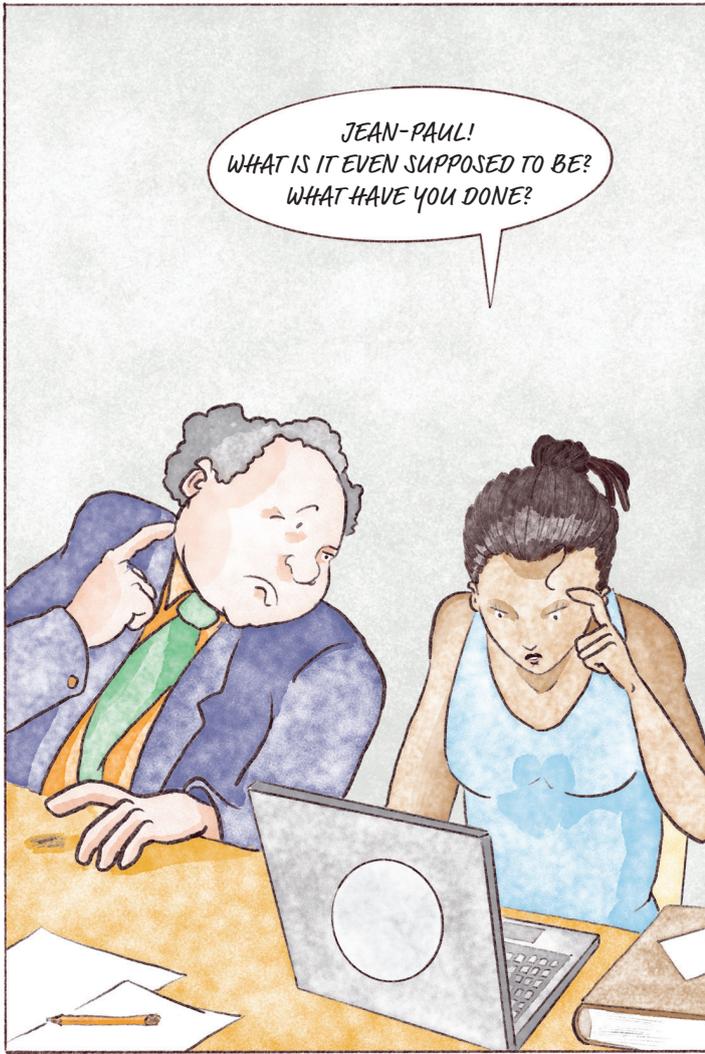


# LET'S TAKE STOCK of **research data** with **Sorella!**









JEAN-PAUL!  
WHAT IS IT EVEN SUPPOSED TO BE?  
WHAT HAVE YOU DONE?



I'VE BEEN TOLD THAT YOU NEED EXCEL  
SPREADSHEETS TO MANAGE THE DATA!  
AND I CAN TELL YOU THAT THESE ARE SOME EXCEL  
SPREADSHEETS! 60 HOURS OF WORK, WITH THE  
SUPPORT OF THREE TRAINEES!

BUT DO YOU REALLY THINK A NORMAL  
PERSON COULD USE THAT?



WELL, THEN! I DO HOPE THEY CAN READ THIS  
BECAUSE THEY'RE REQUESTING IT! OR IS IT  
JUST A WHIM OF THEIRS TO ANNOY EVERYBODY?  
AS IF WE HAD NOTHING BETTER  
TO DO THAN PRODUCE USELESS STUFF  
THAT NOBODY CARES ABOUT!



GROMLMLMLLL...

OK, I SEE... IF WE'RE ASKED TO MANAGE  
OUR PROJECT'S RESEARCH DATA,  
IT'S NOT A USELESS JOB, YOU KNOW!  
LET'S TAKE IT FROM THE TOP TOGETHER,  
SHALL WE?...

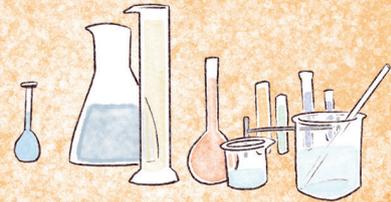
# 1. What are research data?

THERE ARE ACTUALLY MANY DIFFERENT TYPES OF DATA, AND THEY VARY DEPENDING ON THE TOPICS ADDRESSED AND THE DISCIPLINES COVERED. ACCORDING TO CONTEXT, THEY CAN BE:

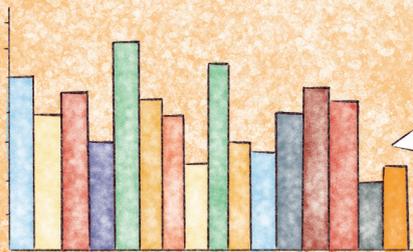
Digital simulation data  
(climate models, economic models, etc.)



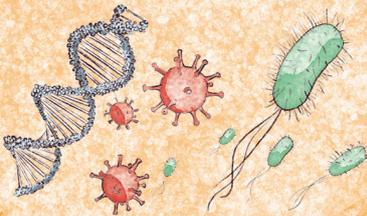
Experimental data  
(biomass weight, peptide sequences, psychology tests, etc.)



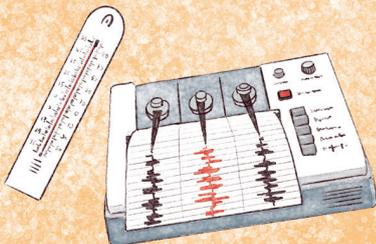
Derived or compiled data  
(compiled databases, text mining, population statistics, etc.)



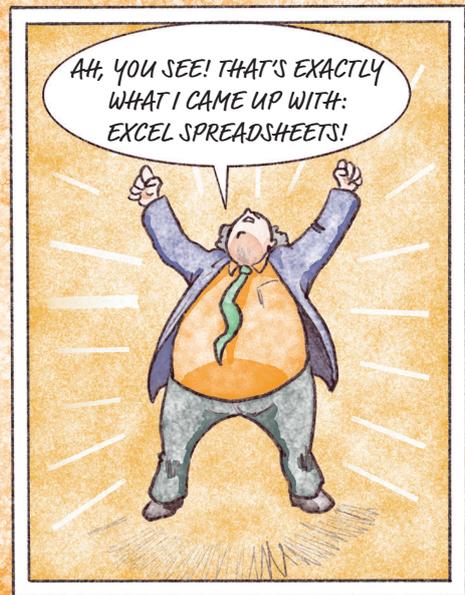
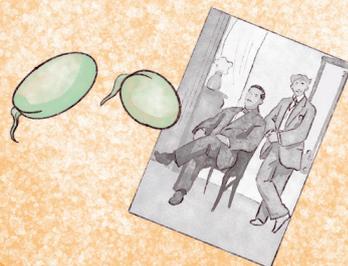
Reference data (text corpuses, TP53 gene sequences, chemical structures, etc.)

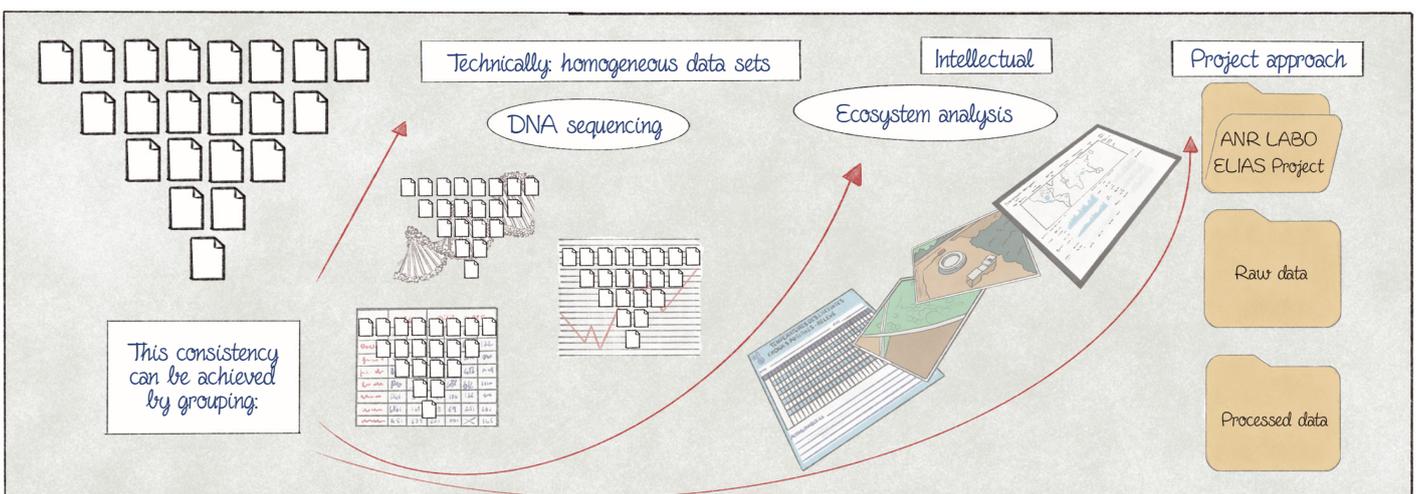
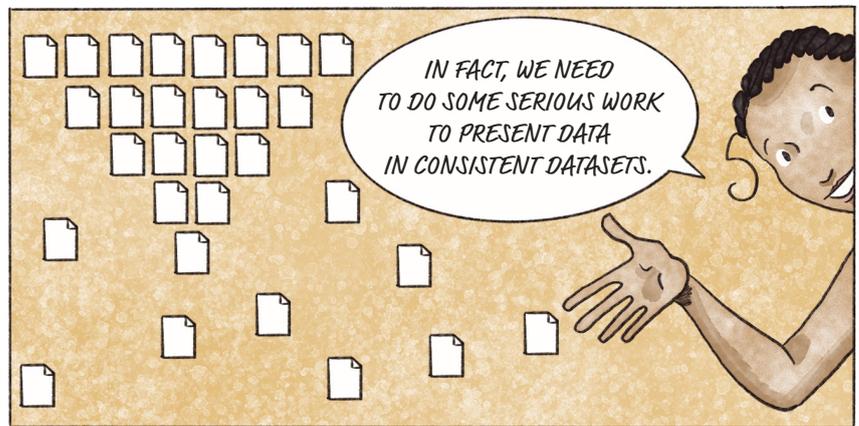
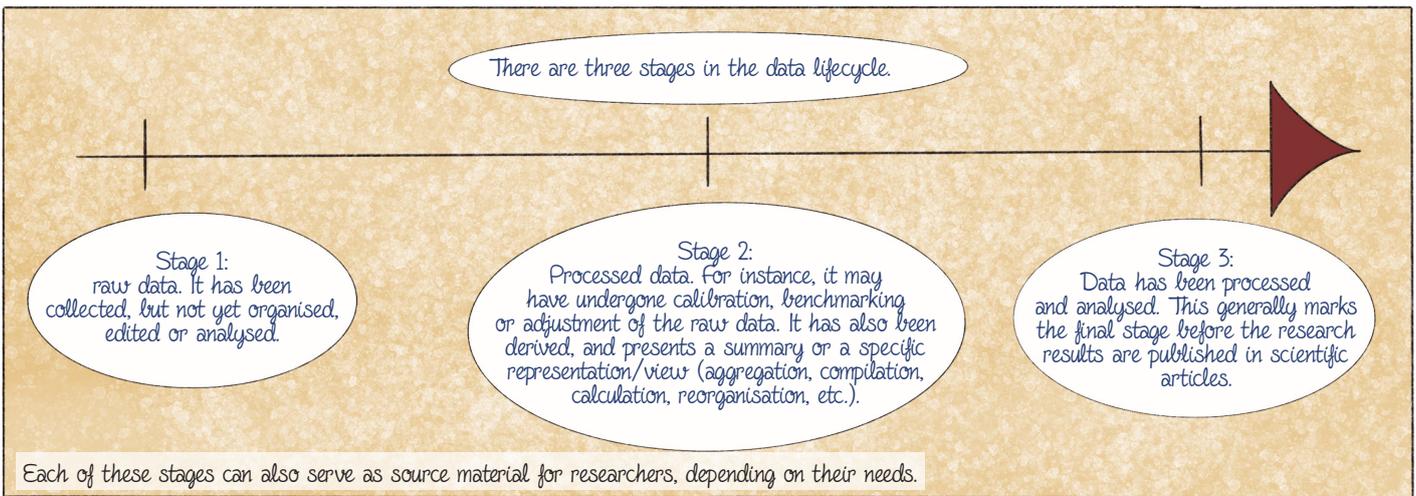
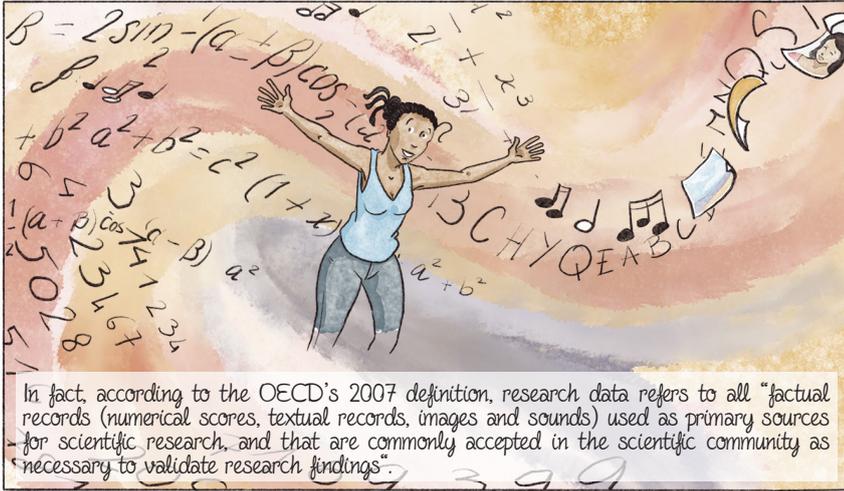


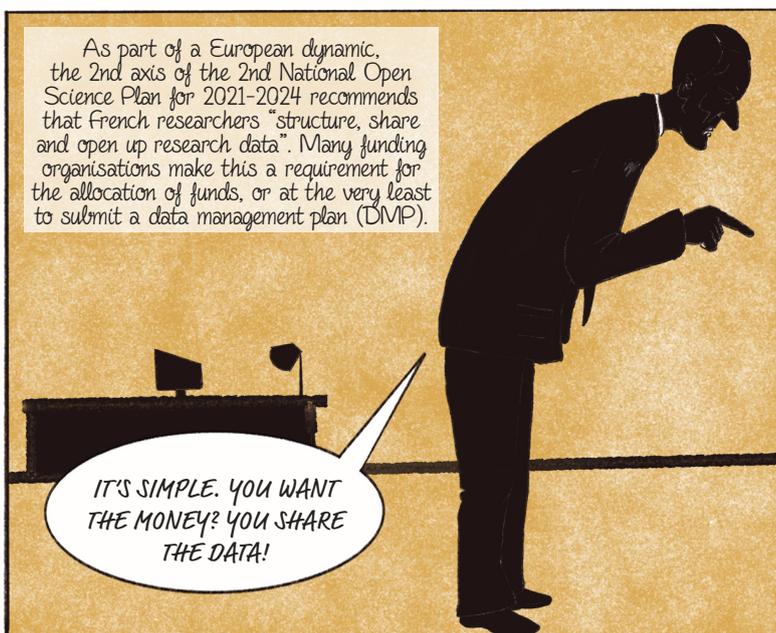
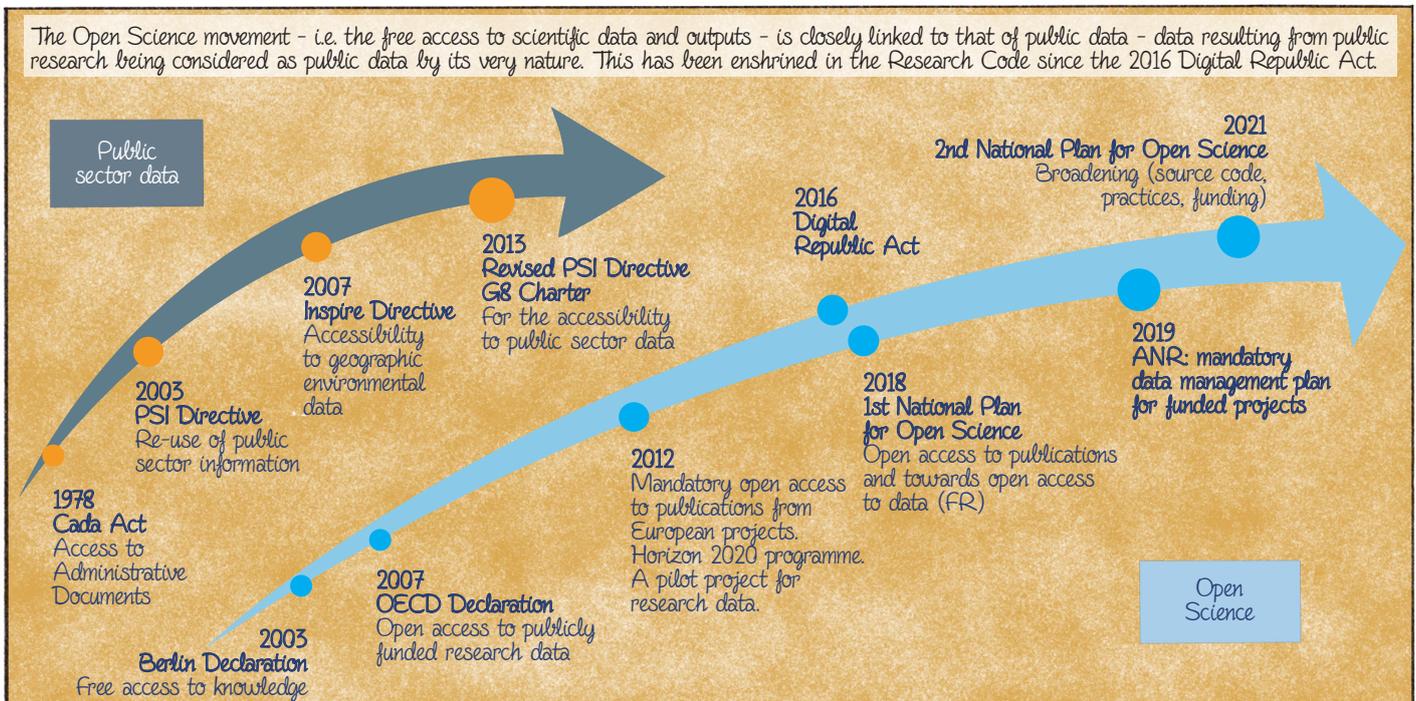
Observation data  
(weather records, seismic readings, images, social surveys, archaeological excavations, etc.)



Archival data (glass plates, photograph stocks, legal texts, plants seeds, etc.)







BECAUSE OPEN SCIENCE IS AT THE CORE OF FIVE KEY PLAYERS' STRATEGIES, ALL OF WHOM HAVE VESTED INTERESTS IN IT.

- Funders and governments
- Scientific publishers
- Research organisations and universities
- Researchers
- Civil society

**FUNDERS AND GOVERNMENTS**  
 Their aim is to promote the re-use of data in order to generate returns on investment and facilitate scientific and technological innovation.

HERE, LOOK. LÉOPOLD DIAGNE CARRIED OUT VIRTUALLY THE SAME STUDY AS I DO. SO DID ELOÏSE SAMPRA, BUT COVERING THE YEARS 2005-2015. I DON'T HAVE ACCESS TO THEIR DATA, OR IT'S JUST UNUSABLE. TOO BAD, I'M GOING TO APPLY FOR FUNDING TO CARRY OUT THE STUDY AGAIN!

ISN'T THAT A BIT SILLY?

**SCIENTIFIC PUBLISHERS**  
 In addition to their duty of meeting the demands of funders, scientific publishers see data sharing as a scientific opportunity to improve the validation of studies, notably by ensuring their reproducibility, and to boost confidence in the presented results.

SOME WEIRD STUFF... RAOUL NEWTON PUBLISHED A STUDY CERTIFYING THAT LEAD COULD FLY UNDER SPECIFIC ATMOSPHERIC CONDITIONS ON EARTH. BUT NOBODY HAS YET MANAGED TO REPRODUCE THE EXPERIMENT WITH THE DATA HE SHARED...

**RESEARCH ORGANISATIONS AND UNIVERSITIES**  
 This is an opportunity for them to promote good ethics and the reproducibility of research within laboratories, by improving resources management and encouraging collaboration between institutions.

HERE IS JEANNINE DUGLAS, A SPECIALIST IN DROSOPHILA REPRODUCTION IN NEW GUINEA, PRESENTING HER NEW PROJECT.

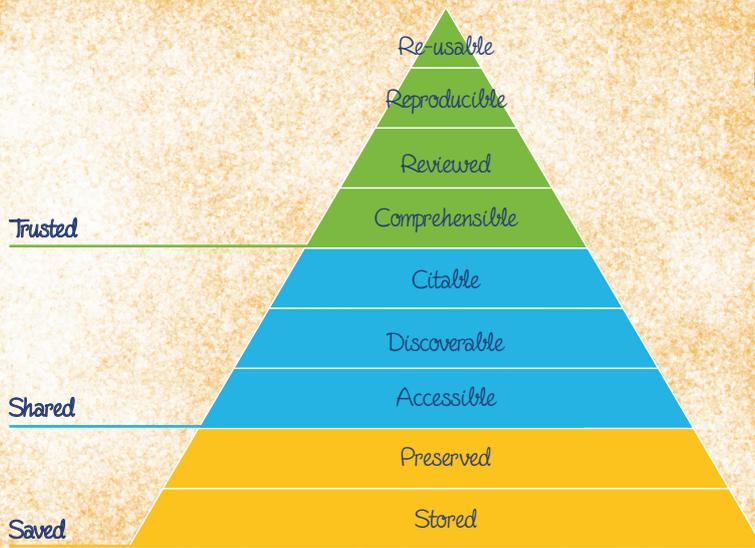
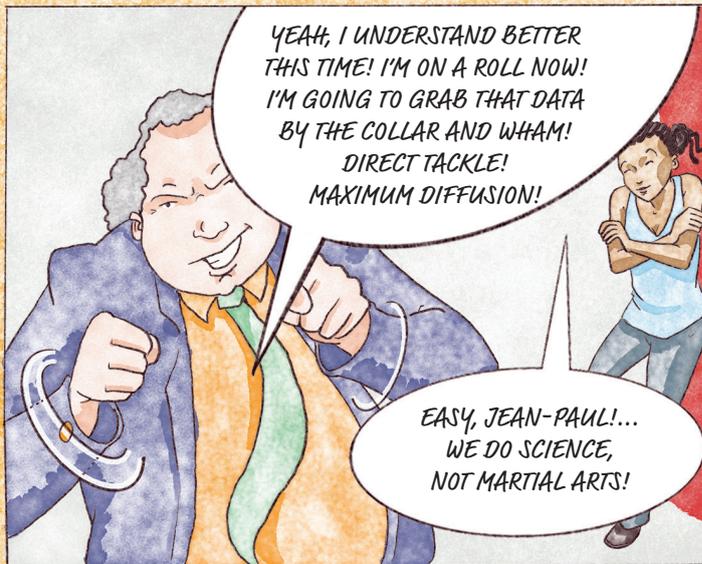
WELL, WE'RE WORKING ON DROSOPHILA DNA. DO YOU THINK WE COULD HAVE HER SEND US THE DATA?

**RESEARCHERS**  
 As for researchers, they learn how to better manage, secure and preserve data, and produce additional academic work that can bring them scientific recognition and credit.

30 YEARS OF RESEARCH! GO FIND SOMETHING NOW IN THIS MESS! I WISH I'D DEALT WITH IT AS IT WENT ALONG...

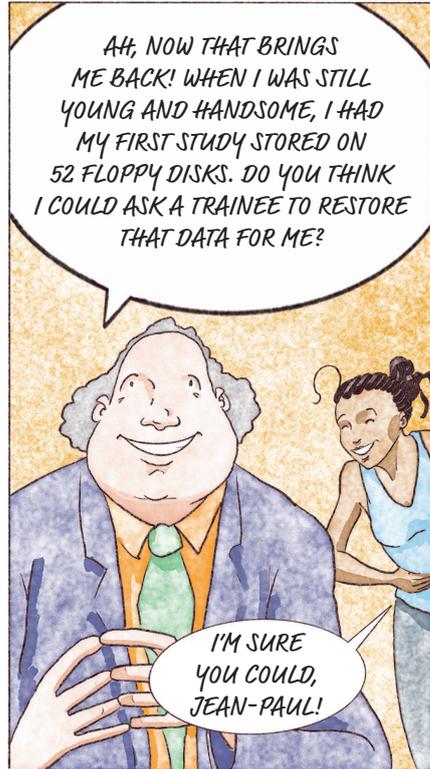
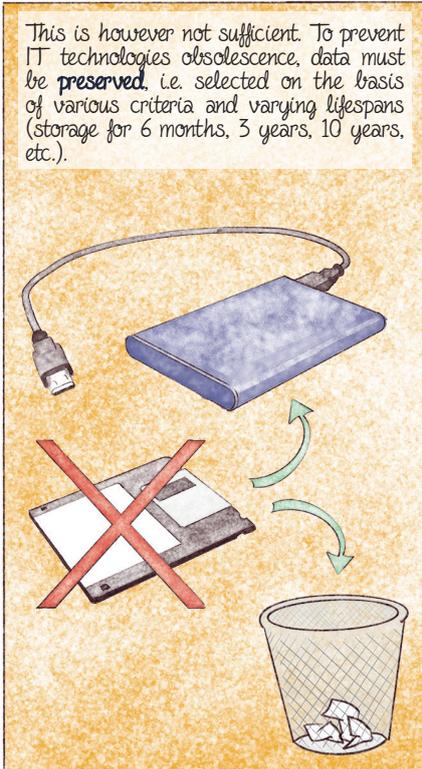
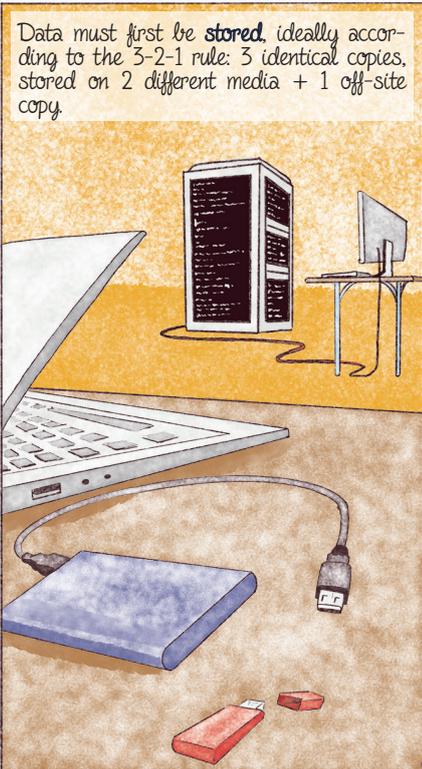
**CIVIL SOCIETY**  
 Last but not least, civil society can benefit from a reliable source of information, encourage innovation and public participation in research.

FANTASTIC! I'VE GOT ALL THE GEO-LOCATION DATA FOR PUBLIC TOILETS IN PARIS: GO FOR A RESTROOMS BOOKING APP!



From A. de Waard. The Mendeley Data management platform: Research data management from a publisher's perspective. (2017) in Danielle Descoteaux, Chiara Farnelli, Marina Soares e Silva, Anita de Waard: Playing Well on the Data FAIRground: Initiatives and Infrastructure in Research Data Management. Data Intelligence 2019; 1 (4): 350-367. doi: [https://doi.org/10.1162/dint\\_a\\_00020](https://doi.org/10.1162/dint_a_00020)

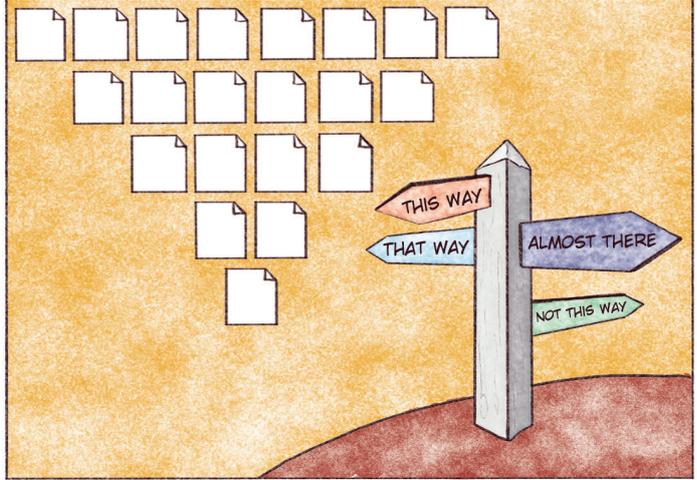
Step 1: Saved data



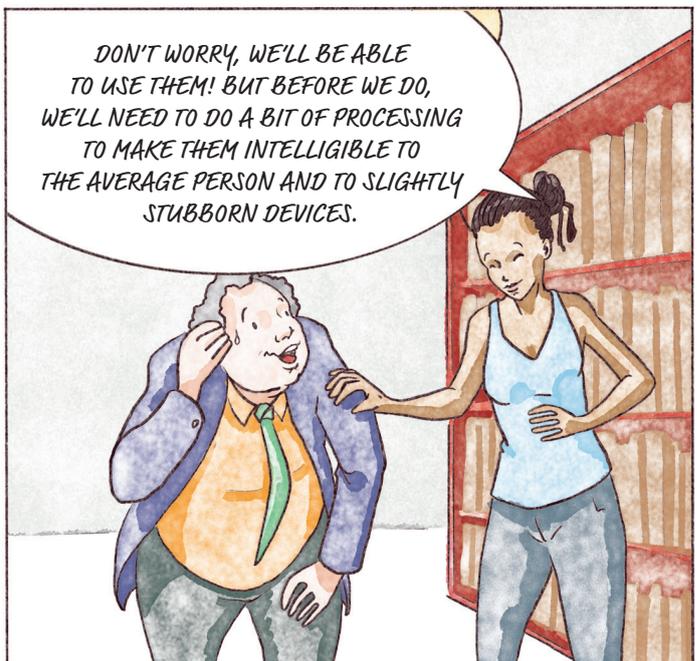
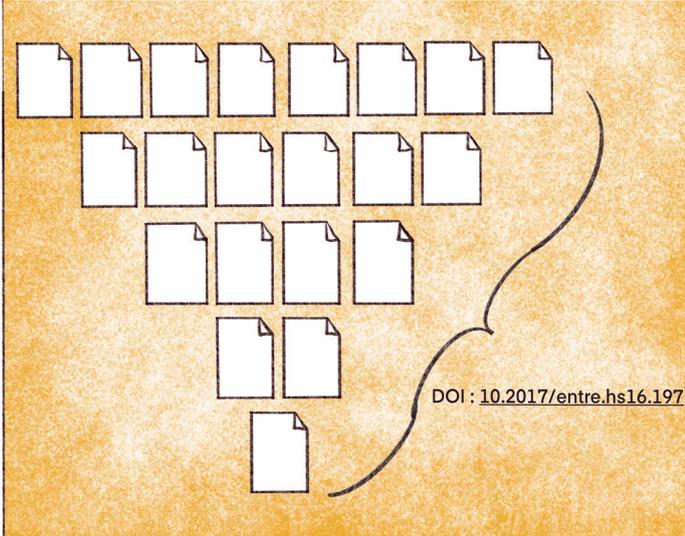
Even when data is stored and preserved, it is not necessarily available to researchers and other devices seeking to interrogate it. At the very least, it should be made **accessible** online.



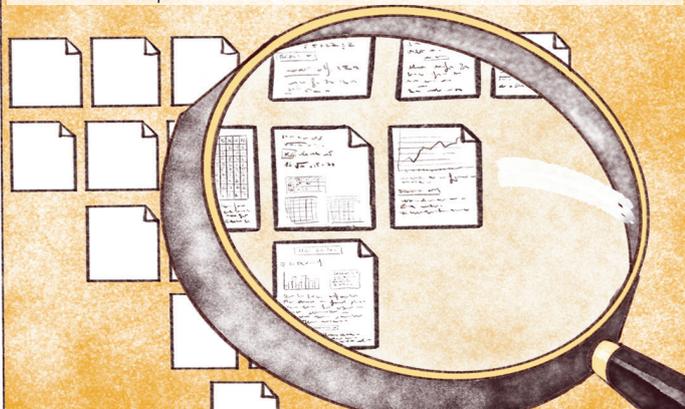
However, even online, research data is not always easy to find by other **researchers**. It is therefore important to make them more widely known - for example by improving the quality of their description.



Furthermore, in order to monitor the re-use of this data and ensure that researchers who produced it receive the scientific credit they deserve, it is advisable to make it **citeable**, for example by assigning it a DOI (Digital Object Identifier) or linking it to a data-paper.



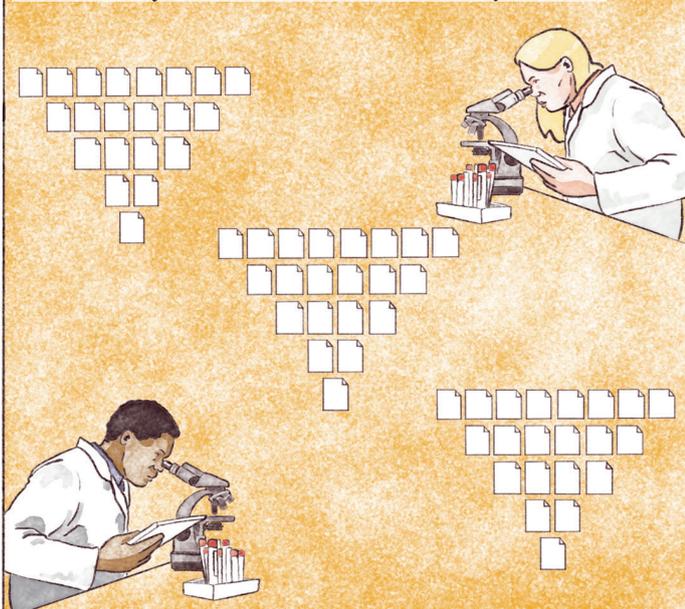
Data that has been collected for internal use is not necessarily **comprehensible** to a third party. It is therefore important to document the data collection: what units of measure were used? What is the context? What abbreviations and parameters were employed? It is also necessary to describe them as accurately as possible - particularly by means of exhaustive and precise metadata.



In order to scientifically **validate** this data, it can also be useful to have it reviewed by peers. Reviewing systems also exist for research data (such as data-papers).



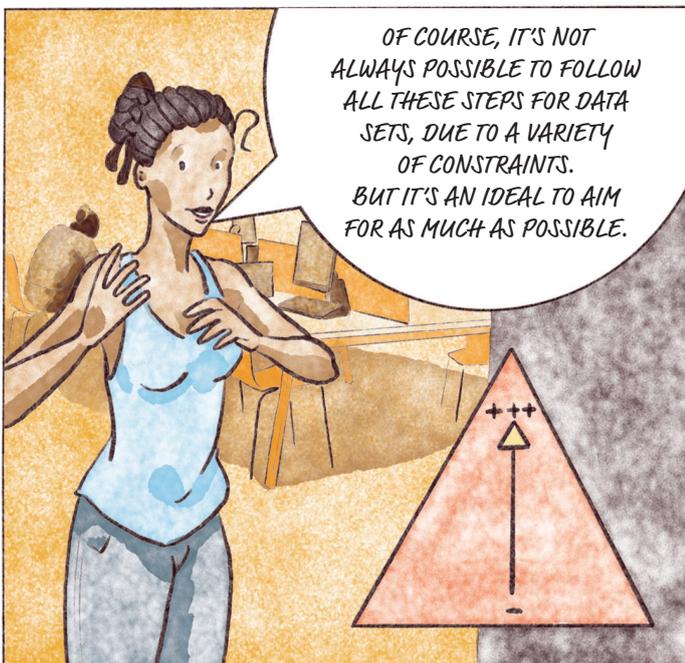
**Reproducibility** of research increases the credibility of results.



Research data that is comprehensible, reliable and reproducible is more likely to be re-used and cited by other researchers. A user licence should therefore be applied to the data, providing a strict framework for **re-use** by other researchers.



OF COURSE, IT'S NOT ALWAYS POSSIBLE TO FOLLOW ALL THESE STEPS FOR DATA SETS, DUE TO A VARIETY OF CONSTRAINTS. BUT IT'S AN IDEAL TO AIM FOR AS MUCH AS POSSIBLE.

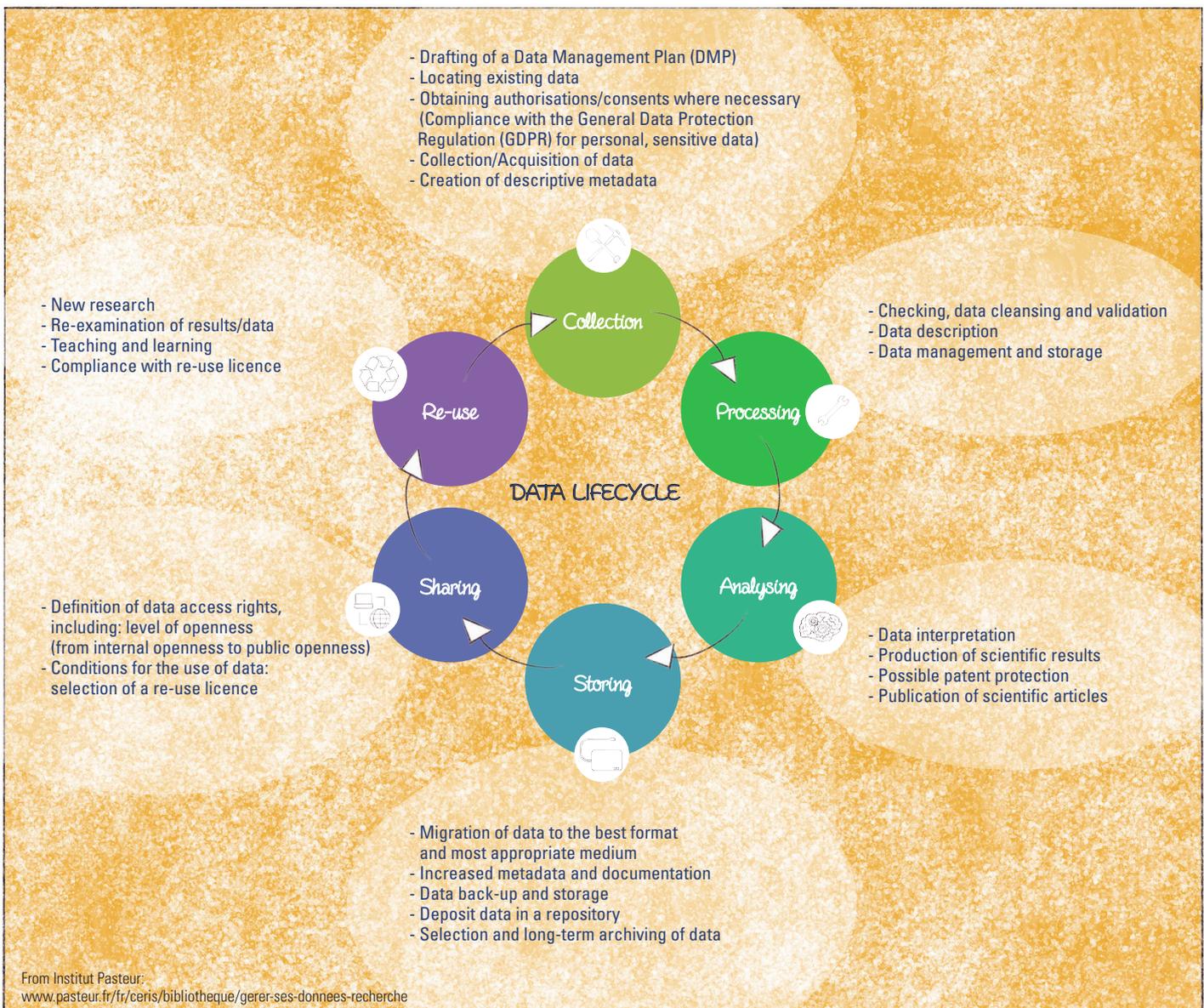
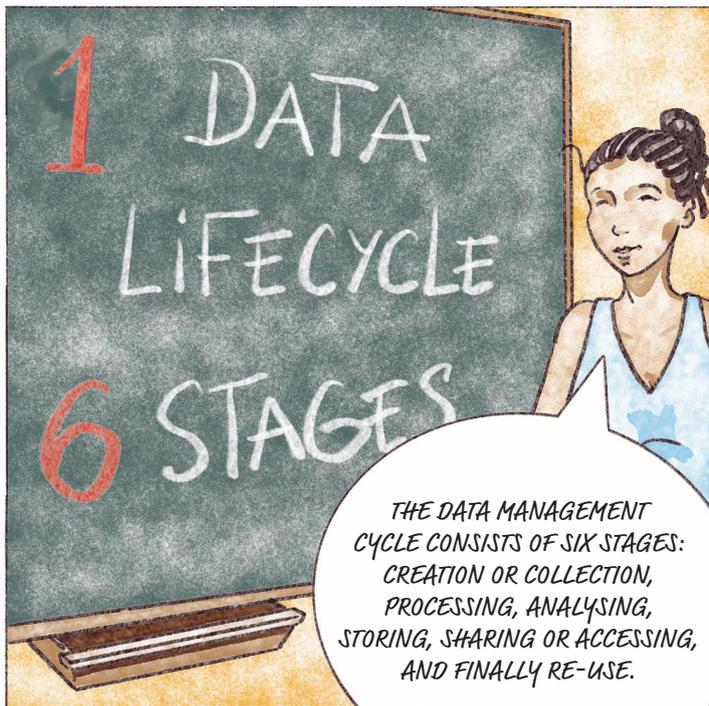


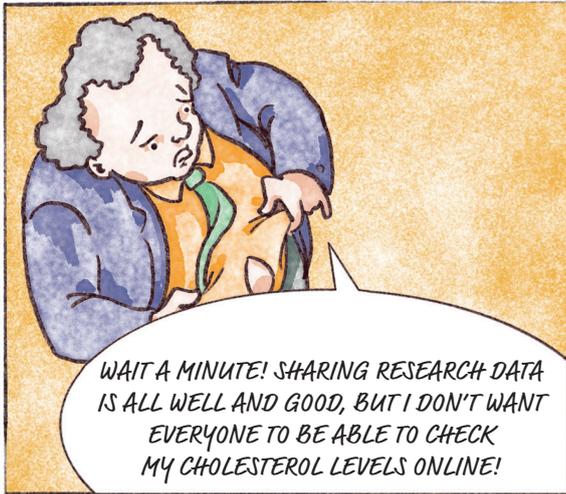
OH DEAR! BUT HOW AM I GOING TO FOLLOW ALL THESE STEPS? I DON'T UNDERSTAND ANY OF IT!



No need to panic! It's now time to put it into practice!

## 2. How to manage and disseminate research data?





LEGAL ISSUES: OPEN OR NON-OPEN DATA?

Data that can be distributed

- Free communication if (cf. Digital Republic Act, Oct. 2016):
- data resulting from a research activity at least semi-financed by public funds
  - not protected by a specific law
  - made public by the researcher or the institution (the institution decides which data will be open, where and under what conditions it will be deposited).
  - Compulsory disclosure of some geographical and environmental data (cf. Inspire convention and Arrhus convention)

Data that can be distributed under specific conditions

- Data presenting risks for the protection of the nation's scientific and technical potential (cf. "protected unit" laboratory)
- Restricted areas: physical and digital access subject to authorisation
- Data protected by copyright and other intellectual property laws
- Personal data (see General Data Protection Regulation (GDPR))

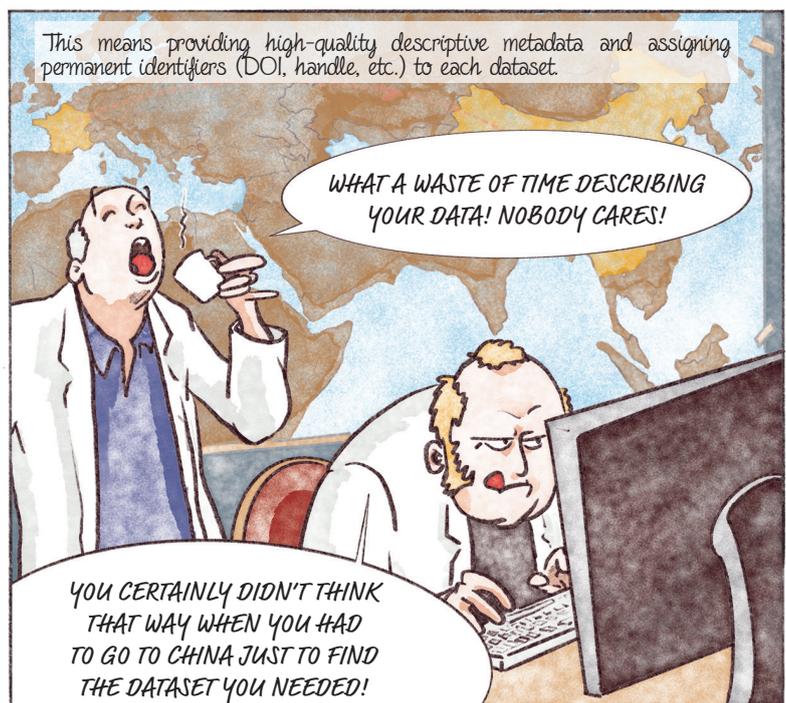
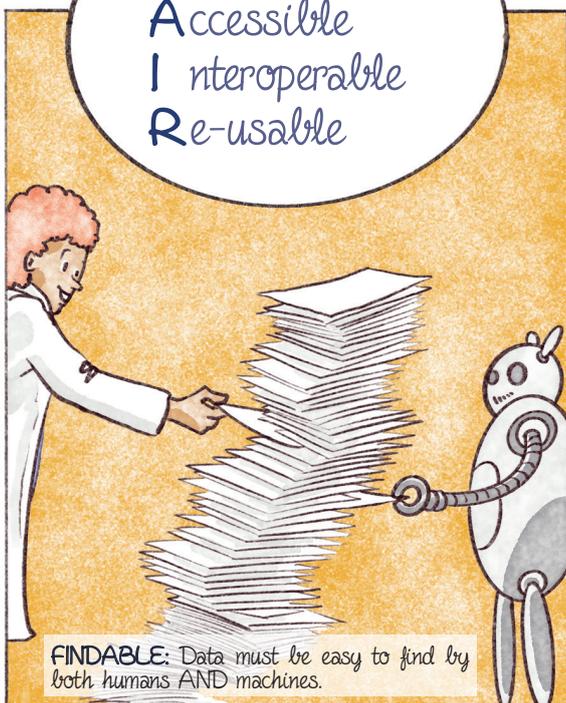
Data prohibited for distribution

- Data presenting risks for the protection of national defense secrets
- Data presenting risks for the security of the State, public safety or the security of the institution
- Professional secrecy or confidentiality (medical secrecy, investigation secrecy, banking and tax secrecy)

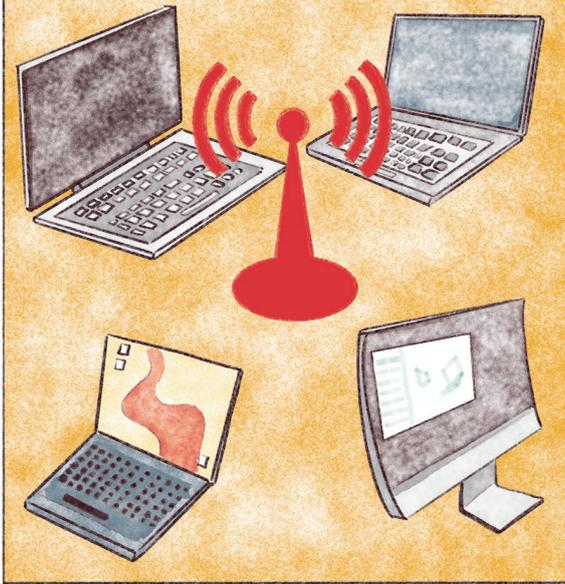
In order to achieve this goal, data sharing must comply with a set of best practices that enable it to be discovered and used by humans - but also by machines. This necessary commitment is summed up by the acronym "FAIR".

From Dominique L'Hostis, From management plan to data-paper, June 2019  
<https://gricad-media.univ-grenoble-alpes.fr/video/plan-gestion-donnees-au-data-paper>

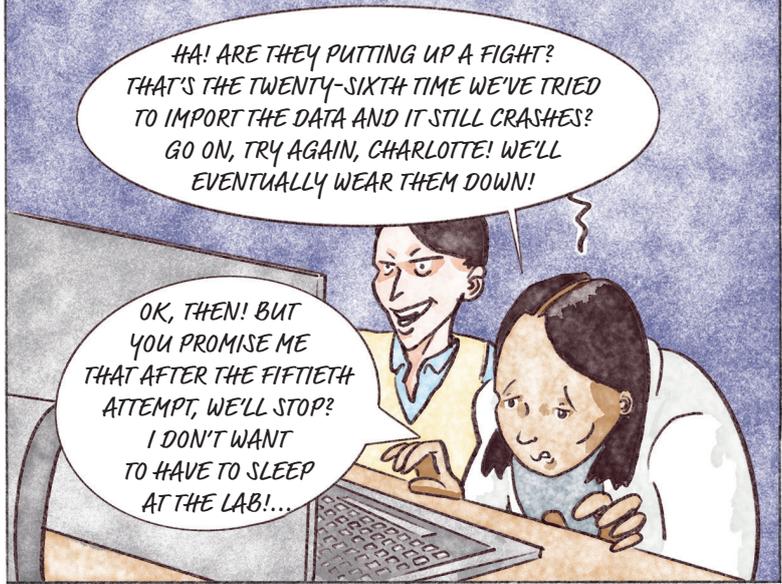
F indable  
 A ccessible  
 I nteroperable  
 R e-usable



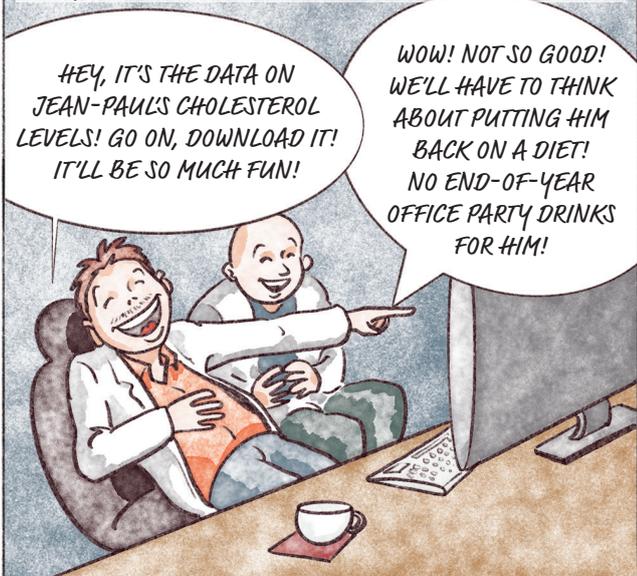
**ACCESSIBLE:** Data and metadata must be stored for the long term, with easy access and/or downloading, specifying the conditions of access and use.



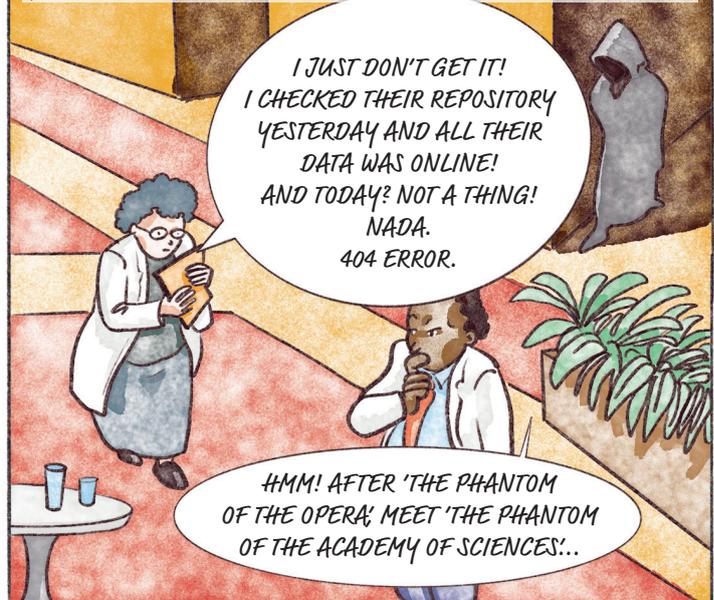
This requires data to be made accessible by its identifiers via a standardised communication protocol (e.g. HTTPS, REST API). It is recommended to mainly use open, free protocols that can be universally implemented.



This also requires protocols to enable authentication and authorisation if necessary, for example to limit or restrict consultation of sensitive, strategic or confidential data to a given type of identified user.



Finally, the metadata must be made accessible even when the data is no longer accessible, which implies establishing long-term archiving protocols.

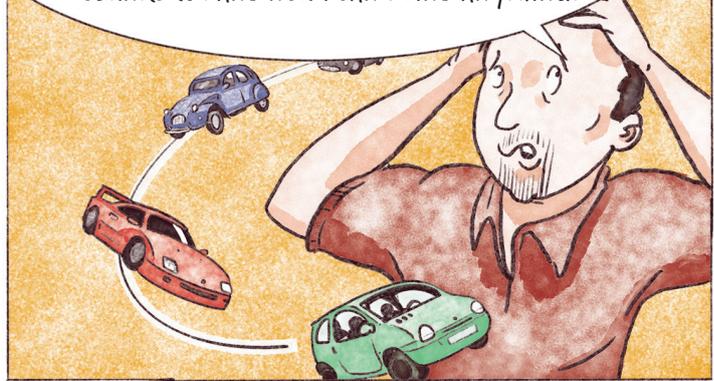


**INTEROPERABLE:** Data and metadata must be downloadable, usable, intelligible and combinable with other data, by humans AND by machines.

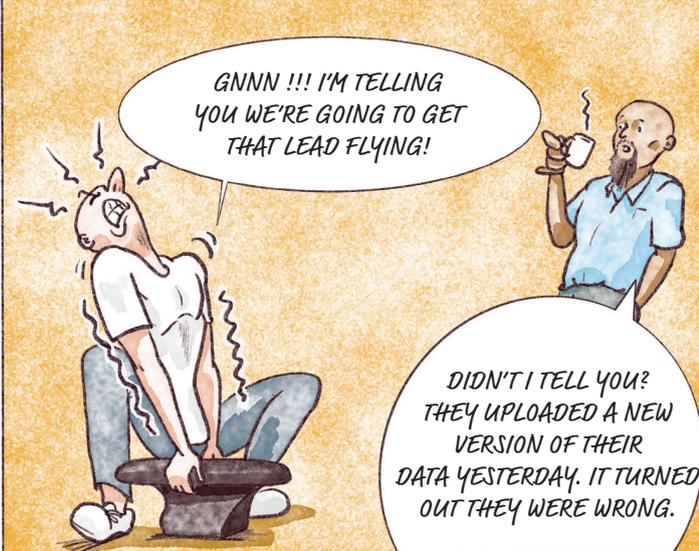


Data and metadata should use a formal, accessible and shared language that is widely applicable to knowledge representation, such as Semantic Web technologies. It is recommended to use standard ontologies and controlled vocabularies.

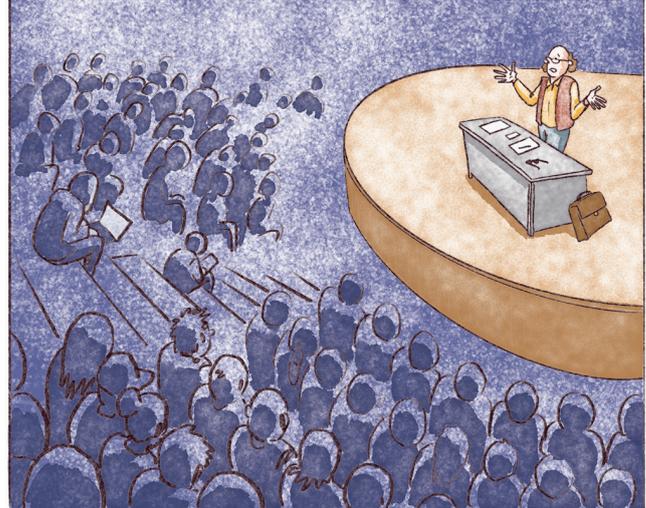
I CAN'T FIND ANY MORE INFORMATION IN THE DATA FROM THE "BEETLE" LABORATORY. TO DESCRIBE THEIR STUDY OF CARS, THEY USED THE WORDS "AUTOMOBILE"; "VEHICLE"; "RIDE" AND "CHARIOT". AT SOME POINT THEY EVEN REFERRED TO THEM AS "JALOPIES" AND "CLUNKERS"! AND NOW I CAN'T FIND ANYTHING!



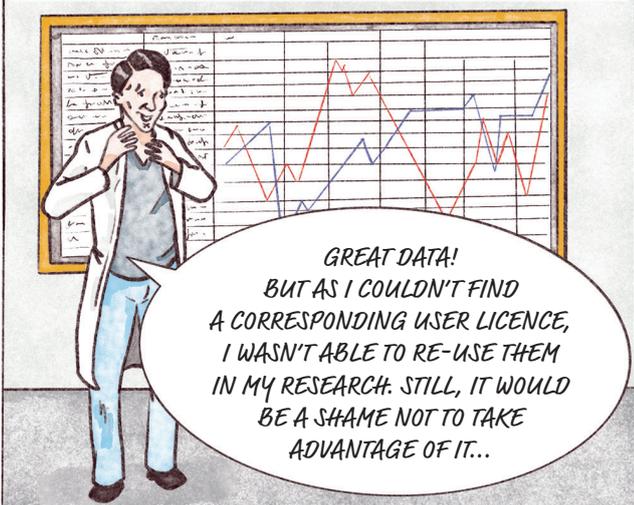
Data and metadata may include links to other (meta)data, previous or more recent versions, additional data or articles citing the data.

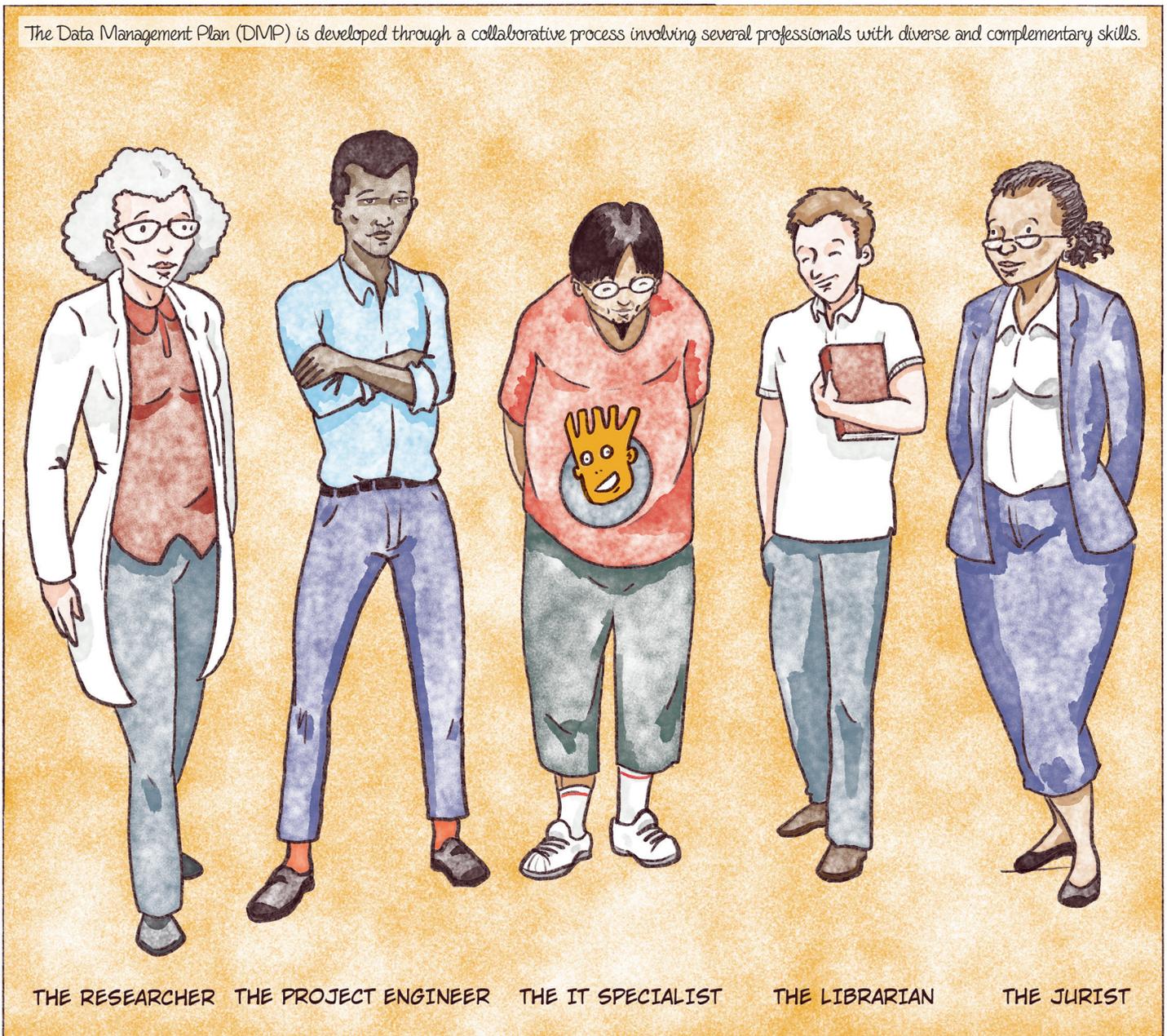
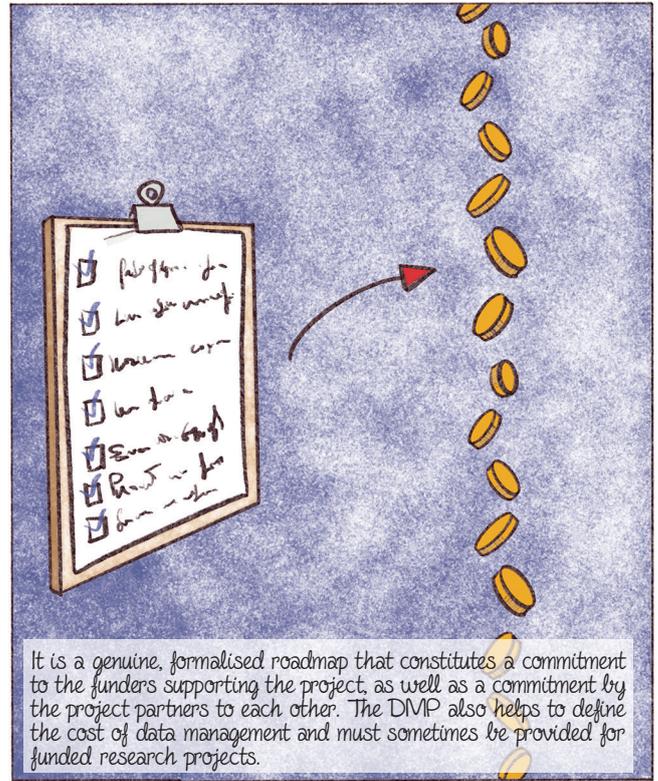
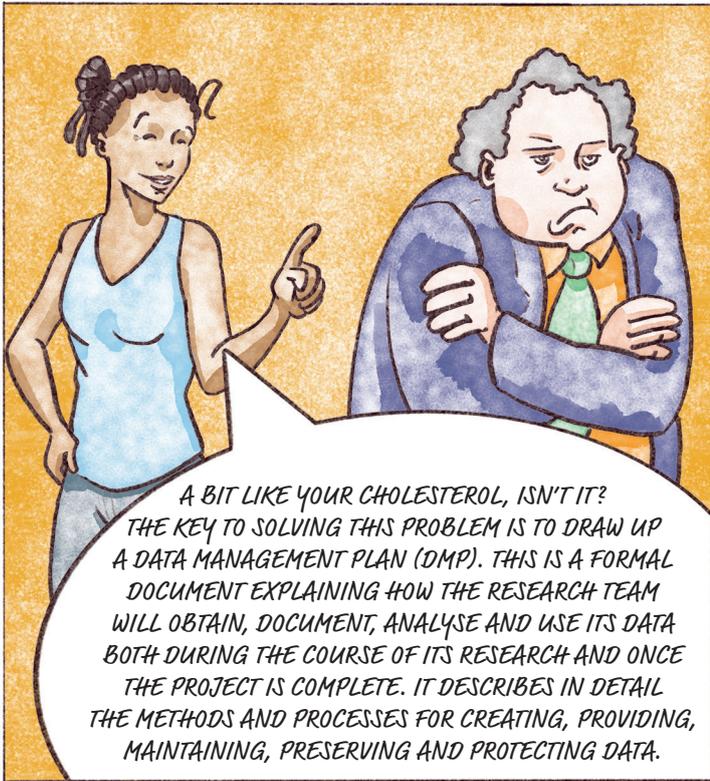


**RE-USABLE:** Data and metadata must include characteristics that make the data re-usable for future research or for other purposes (teaching, innovation, replication and scientific transparency).



To that end, providing datasets with an explicit and accessible user licence and linking them to their provenance in compliance with the standards of the specified communities will make it easier to re-use them later on.





**THE RESEARCHER**

LET'S GO! DOWN TO THE MINE!  
I'M SUPPLYING THE RAW MATERIAL  
FOR THE PROJECT!

The researcher is in charge of collecting, describing and breaking down the data into consistent sets.

**THE PROJECT ENGINEER**

LEFT! RIGHT!  
BACKWARDS! STRAIGHT  
AHEAD! COME ON,  
GUYS! WE'RE GONNA  
MAKE IT!

When a project engineer is appointed to the project, he or she becomes responsible for coordinating the actions performed around the data.

**THE IT SPECIALIST**

I SAID "NO", JEAN-PAUL!  
WE'RE NOT UPLOADING  
YOUR EXCEL FILE  
TO A CLOUD!

The IT specialist is the key contact when it comes to data storage and security, infrastructure and cost aspects.

**THE LIBRARIAN**

ARE YOU SURE YOU WANT  
TO SHARE YOUR DATA ON THE HISTORY  
OF THE BEETLE FOR THE DURATION  
OF THE PARIS MOTOR SHOW ONLY?  
WOULDN'T THAT BE A BIT BRIEF?

The Scientific and Technical Information specialist - whether documentalist or archivist - helps the researcher to select the data, define the storage timeframes and the appropriate technical solutions for sharing the data.

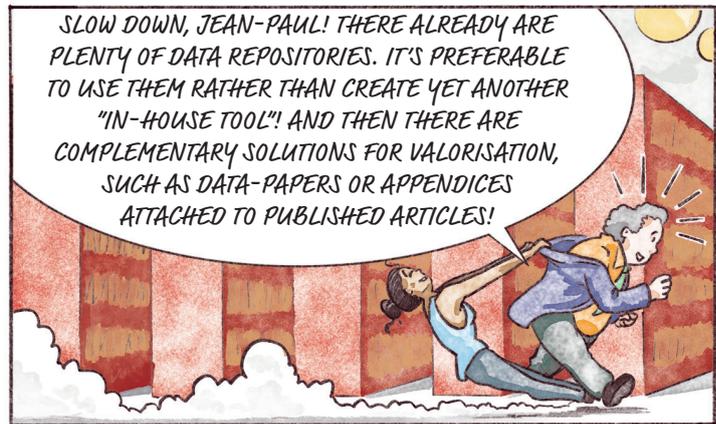
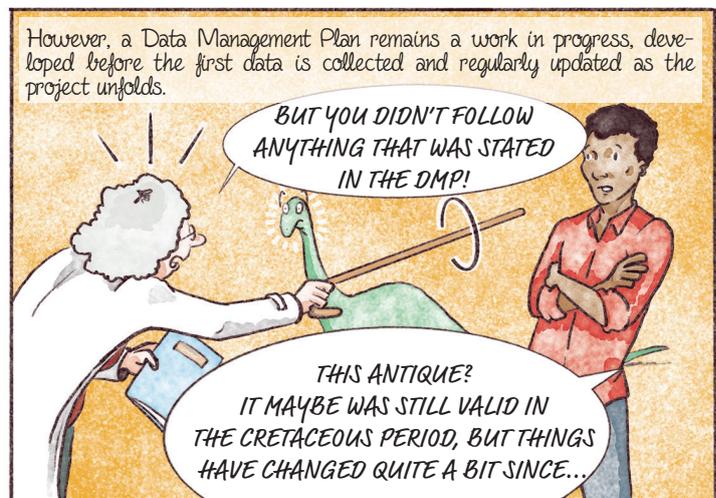
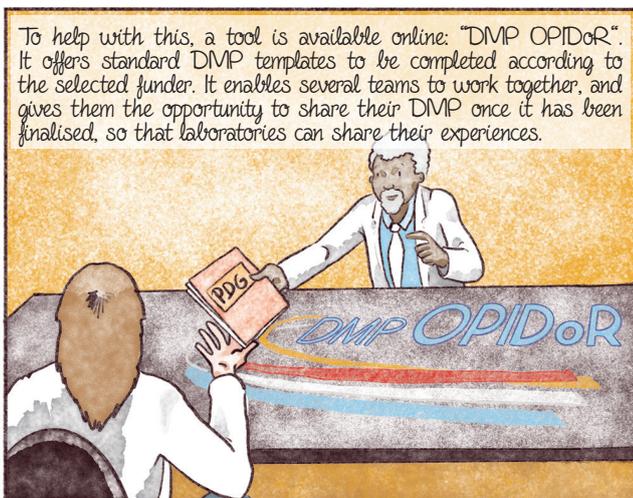
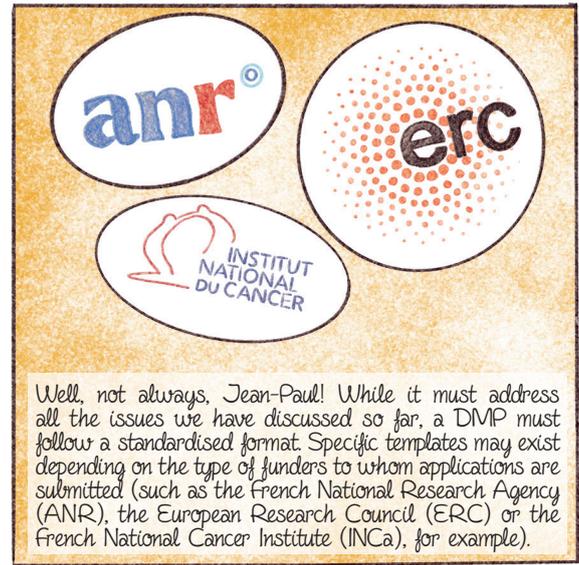
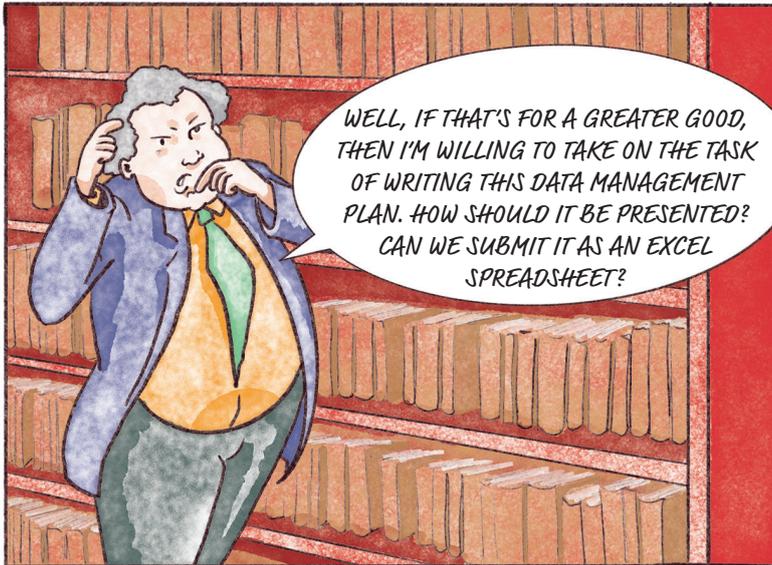
**THE JURIST**

ENOUGH! I SAID WE WERE  
DONE WITH JEAN-PAUL'S  
CHOLESTEROL LEVELS!

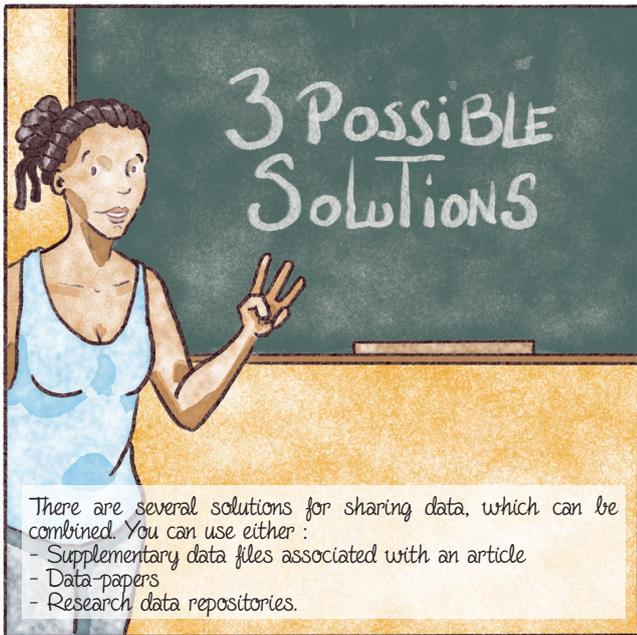
The jurist and the DPO (Data Protection Officer) advise on data anonymisation procedures, on licences to be used, and on access permissions to be granted to various user groups (open, restricted or embargoed access). They also provide advice on protecting data, in order to ensure that it can be economically leveraged through patents or other means.

**OTHER PROFESSIONALS**

This list of professionals who may be involved in drafting a data management plan (DMP) is obviously not exhaustive. Many other skills can be brought on board.



### 3. Sharing research data: which tools to use?



There are several solutions for sharing data, which can be combined. You can use either :

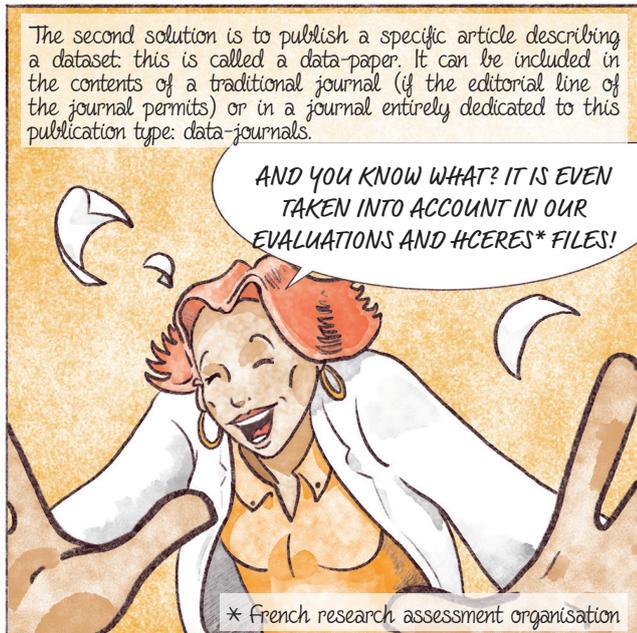
- Supplementary data files associated with an article
- Data-papers
- Research data repositories.



THE FIRST SOLUTION IS TO PUBLISH YOUR ARTICLE IN A JOURNAL WITH ARCHIVED DATA APPENDED. THIS CAN TAKE THE FORM OF...

EXCEL SPREADSHEETS!

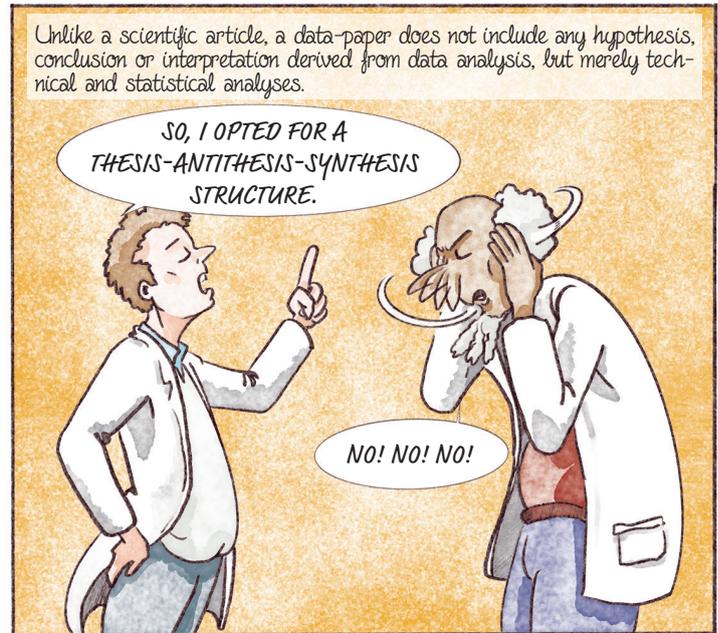
The second solution is to publish a specific article describing a dataset: this is called a data-paper. It can be included in the contents of a traditional journal (if the editorial line of the journal permits) or in a journal entirely dedicated to this publication type: data-journals.



AND YOU KNOW WHAT? IT IS EVEN TAKEN INTO ACCOUNT IN OUR EVALUATIONS AND #CERES\* FILES!

\* French research assessment organisation

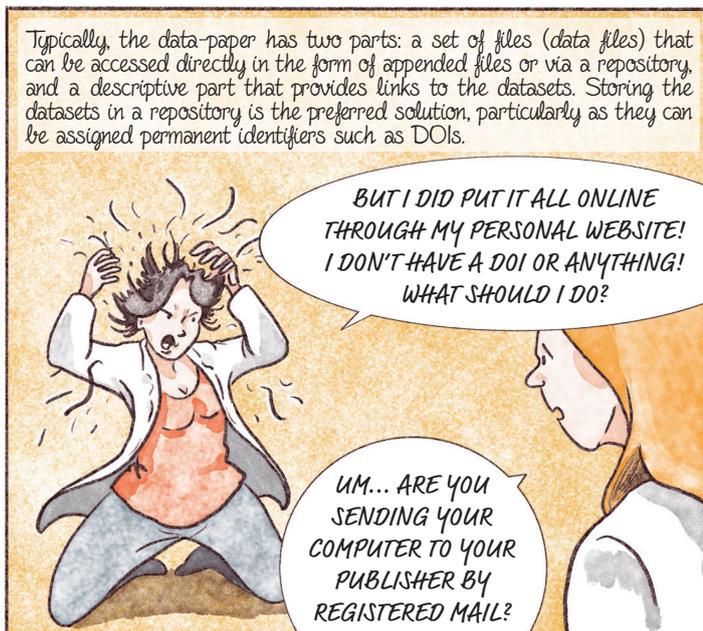
Unlike a scientific article, a data-paper does not include any hypothesis, conclusion or interpretation derived from data analysis, but merely technical and statistical analyses.



SO, I OPTED FOR A THESIS-ANTITHESIS-SYNTHESIS STRUCTURE.

NO! NO! NO!

Typically, the data-paper has two parts: a set of files (data files) that can be accessed directly in the form of appended files or via a repository, and a descriptive part that provides links to the datasets. Storing the datasets in a repository is the preferred solution, particularly as they can be assigned permanent identifiers such as DOIs.

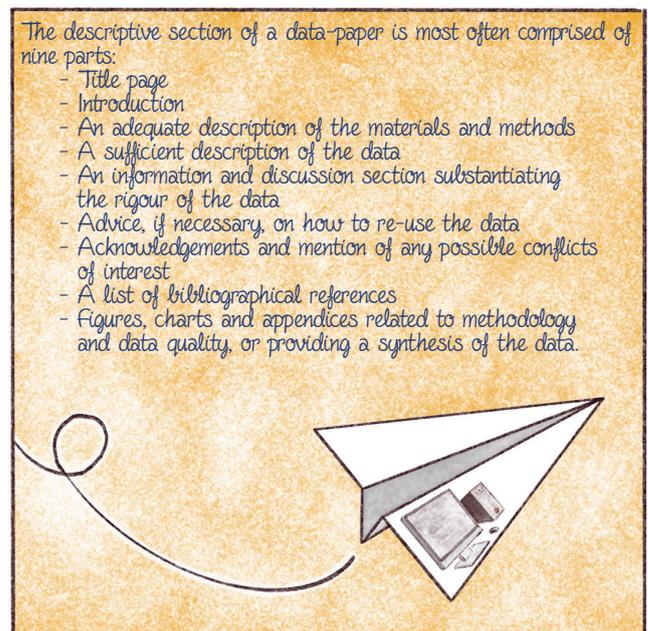


BUT I DID PUT IT ALL ONLINE THROUGH MY PERSONAL WEBSITE! I DON'T HAVE A DOI OR ANYTHING! WHAT SHOULD I DO?

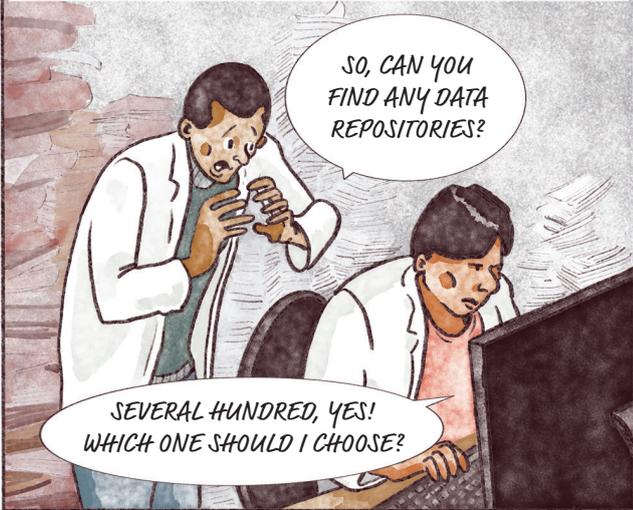
UM... ARE YOU SENDING YOUR COMPUTER TO YOUR PUBLISHER BY REGISTERED MAIL?

The descriptive section of a data-paper is most often comprised of nine parts:

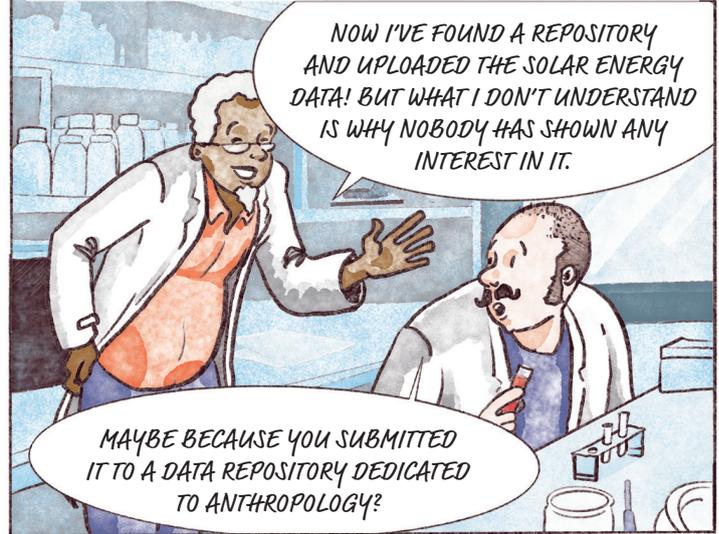
- Title page
- Introduction
- An adequate description of the materials and methods
- A sufficient description of the data
- An information and discussion section substantiating the rigour of the data
- Advice, if necessary, on how to re-use the data
- Acknowledgements and mention of any possible conflicts of interest
- A list of bibliographical references
- Figures, charts and appendices related to methodology and data quality, or providing a synthesis of the data.



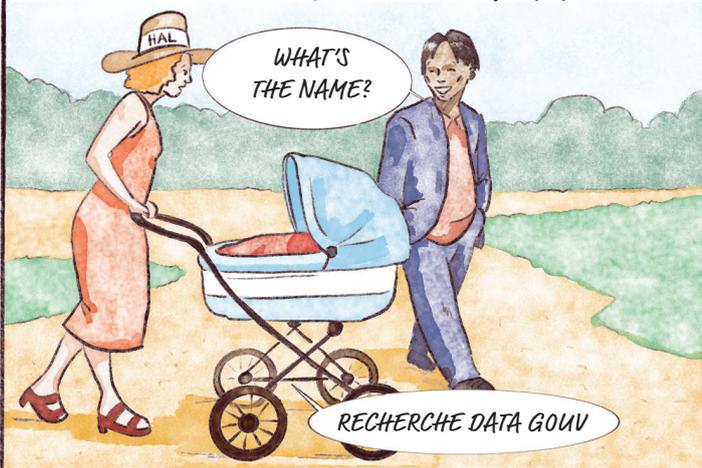
The third solution, which is also the most recommended, is to host your datasets on a dedicated data repository. A data repository is a reservoir of research data, raw or derived, that can be retrieved and re-used thanks to a metadata description.



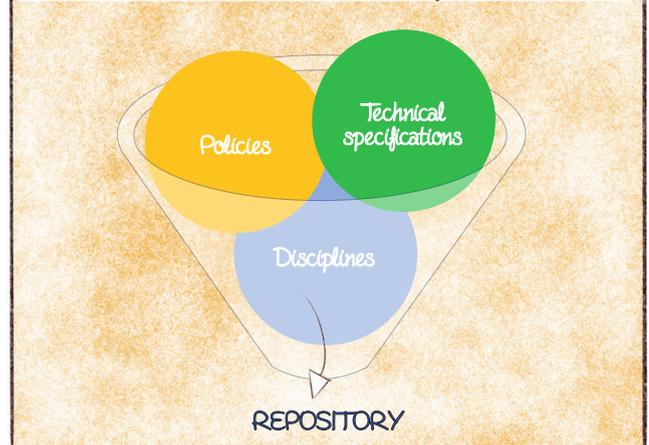
That's right! There are several thousand data repositories around the world. They fall into three main categories: disciplinary, generalist and institutional.



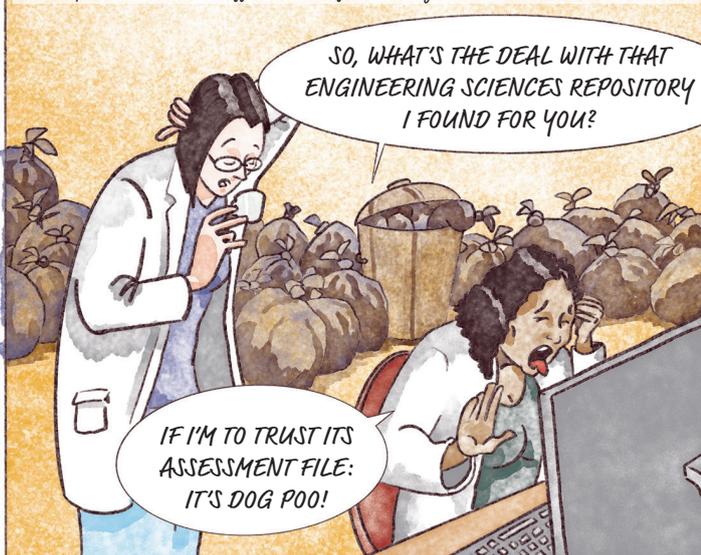
Since the end of 2021, France has been operating a national repository for research data, funded by the Ministry of Higher Education and Research, called Recherche Data Govv: <https://recherche.data.gouv.fr/fr>



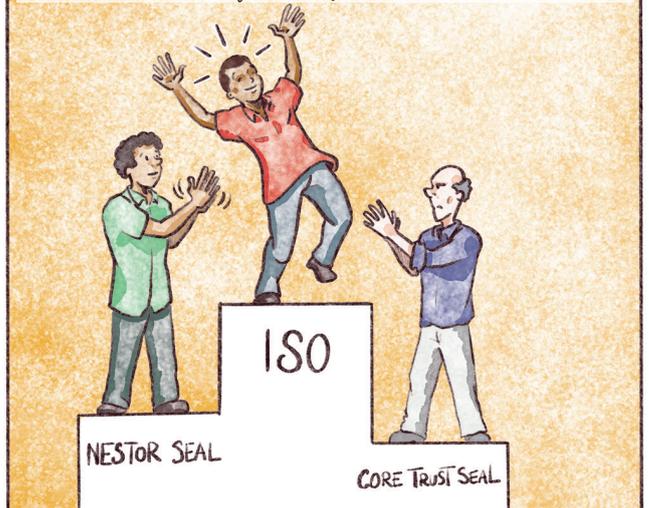
As you can see, selecting the right repository for your data is essential. This involves three main criteria: the technical specifications of the repository, whether it meets the policy requirements of funders or publishers, and the disciplinary aspect.

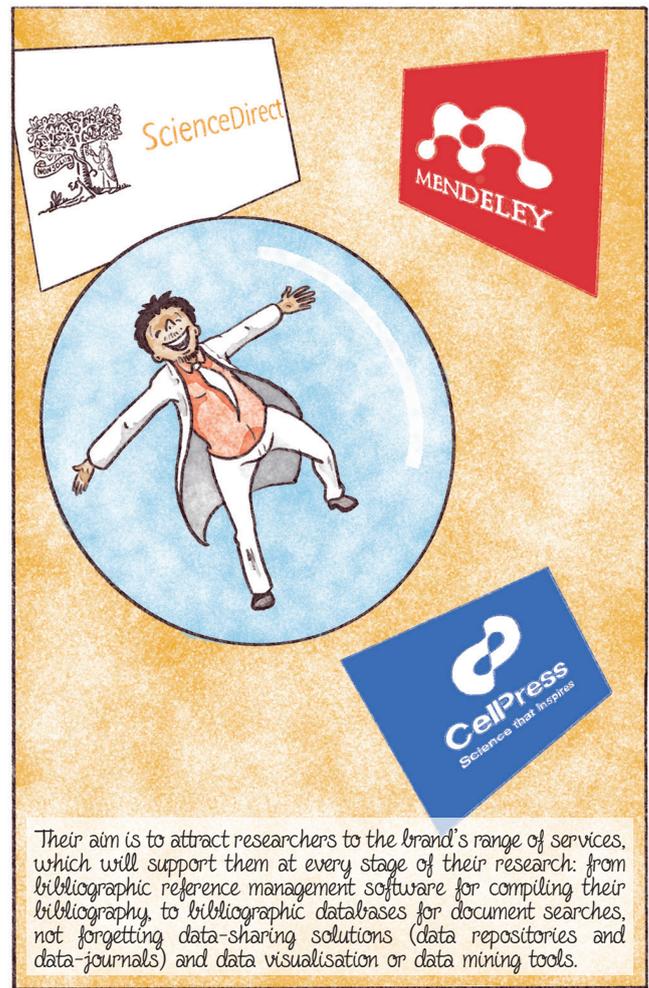
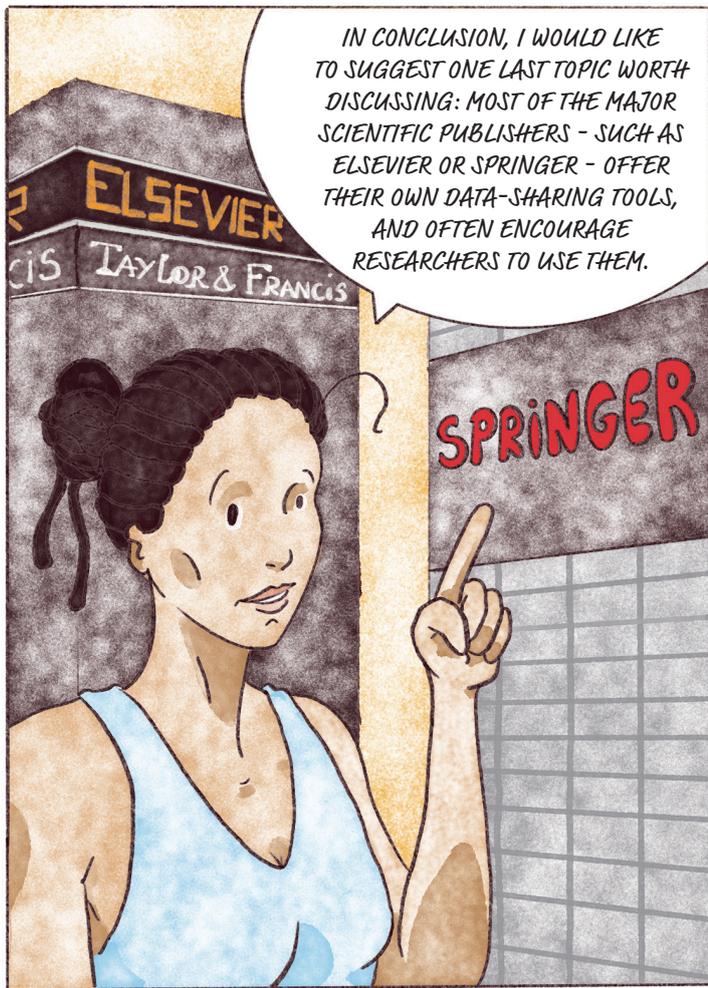


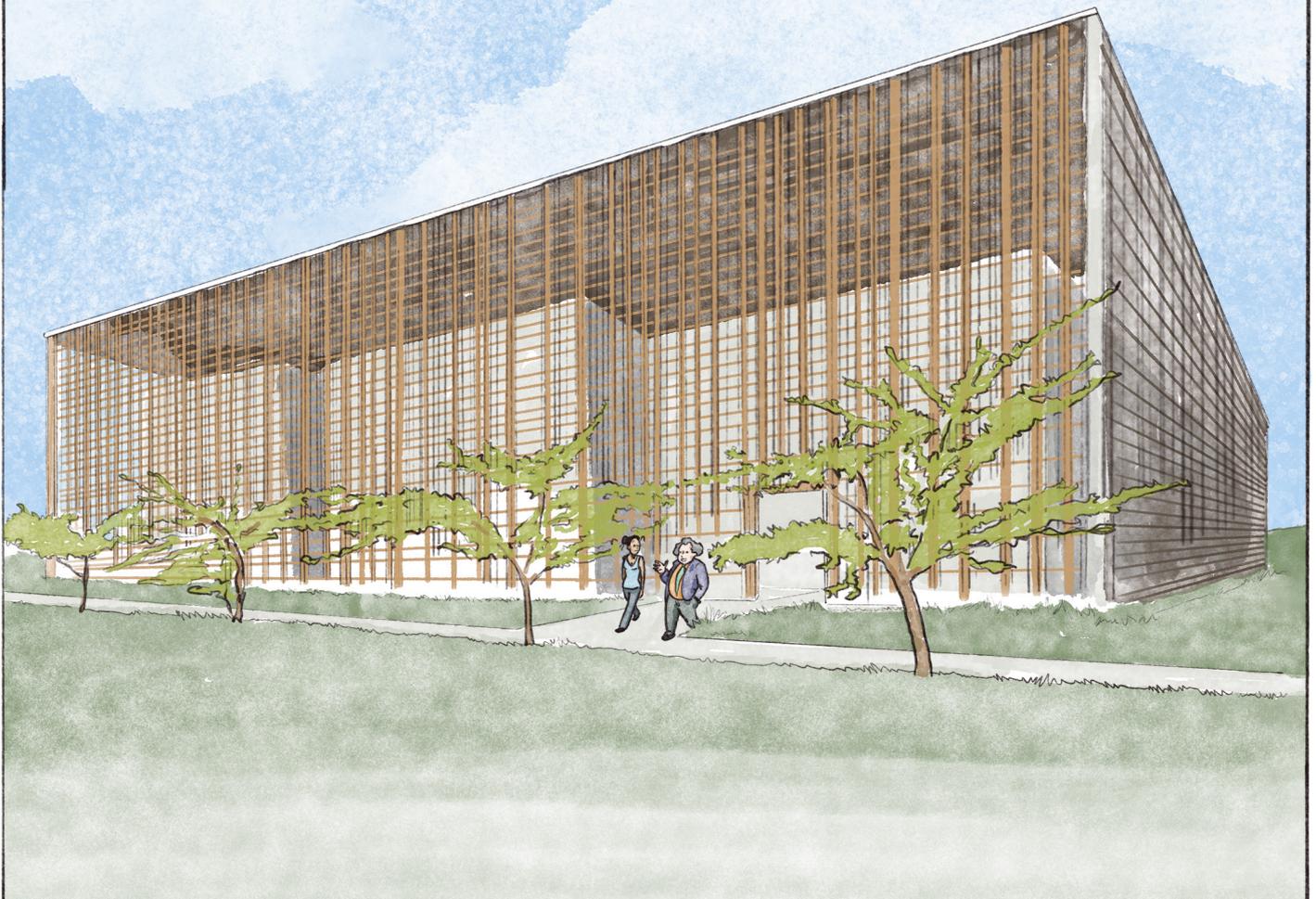
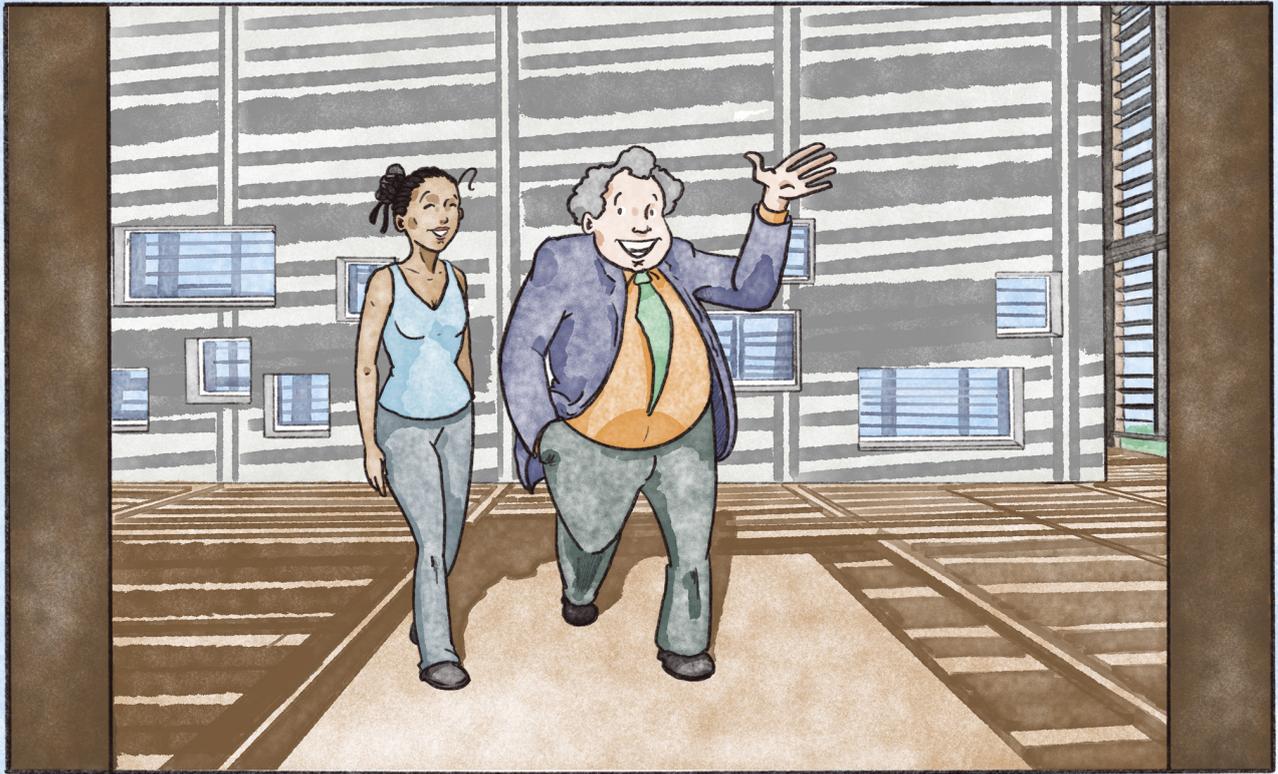
Regarding the technical specifications of the repositories, a certification system has been introduced to help researchers find their way around. These certifications are based on the evaluation of the repository by an independent jury on a declaratory basis. Access to the assessment file is made public to allow effective verification by users.



Certification attests to the repository's compliance with the "FAIR" principles. There are currently three levels of certification: 'CoreTrustSeal' is obtained after meeting 16 criteria (134 centres were certified worldwide in 2023). 'Nestor Seal' requires a commitment to 34 criteria (5 centres certified worldwide in 2023). Finally, to obtain ISO+, you have to comply with around a hundred criteria (only one certified centre worldwide in 2023)!







### **BIBLIOGRAPHICAL REFERENCES:**

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Callisto Formation. Fondation UNIT.  
[https://callisto-formation.fr/?theme=boostplus\\_c06&redirect=0](https://callisto-formation.fr/?theme=boostplus_c06&redirect=0)

CoopIST : délégation à la formation scientifique et technique, CIRAD.  
Gérer les données de la recherche.  
<https://doi.org/10.18167/COPIST/0005>

DoRANum – Données de la recherche : Apprentissage Numérique.  
<https://doranum.fr/>

Ouverture des données de recherche – Guide d’analyse  
du cadre juridique en France – V2. (2017). Ouvrir la science !  
Comité pour la science ouverte.  
[www.ouvirlascience.fr/ouverture-des-donnees-de-recherche-guide-danalyse-du-cadrejuridique-en-france-v2](http://www.ouvirlascience.fr/ouverture-des-donnees-de-recherche-guide-danalyse-du-cadrejuridique-en-france-v2)

Partager les données liées aux publications scientifiques –  
Guide pour les chercheurs. (2022). Ouvrir la Science !  
Comité pour la science ouverte.  
[www.ouvirlascience.fr/partager-les-donnees-liees-aux-publications-scientifiques-guidepour-les-chercheurs](http://www.ouvirlascience.fr/partager-les-donnees-liees-aux-publications-scientifiques-guidepour-les-chercheurs)

**AUTHOR**

*Marie Latour*, deputy director of the library, University of French Guiana

**SCIENTIFIC SUPERVISOR**

*Annaïg Mahé*, lecturer at URFIST Paris

**ARTWORK**

*Olivier Copin*

**GRAPHIC DESIGN**

*Bénédicte Sauvage* (BCOM)

**ENGLISH TRANSLATOR**

*Stéphane Berland*

**SCIENTIFIC PROOFREADING**

*Cyril Heude* (data librarian at SciencePo Paris),  
*Romain Féret* (director of Média Normandie),  
*Amélie Barrio* (Co-director of URFIST Occitanie)

Project co-financed by URFIST Paris  
(Ecole nationale des Chartes - PSL)

