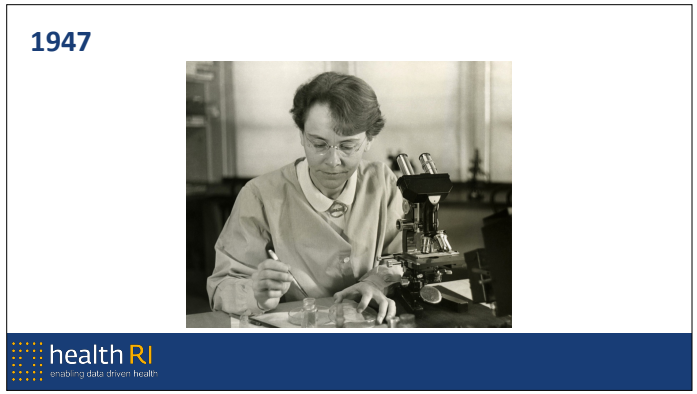
Announced Tiltle: Research DMP: where to start.

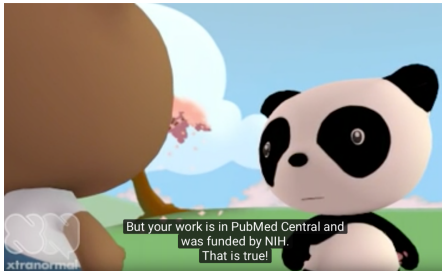
I will take you along a journey of research first taking a look at “data”, then via FAIR Data and FAIR Data Management, and ending up with FAIR Data Management Planning.



This is the first of two pictures I want to show you of researchers. This is
* Barbara McClintock, Nobel prize in 1983 Physiology or Medicine, at work at her microscope in 1947

There are notable differences with this picture of researchers at a microscope taken 70 years later.
* Computers and digital data are how we process information now
* Algorithms interpret data
* Volumes grow immensely
* Types grow immensely: genomics, proteomics, metabolomics
* Datasets become important output **by themselves**

Dystopia (movie to watch)



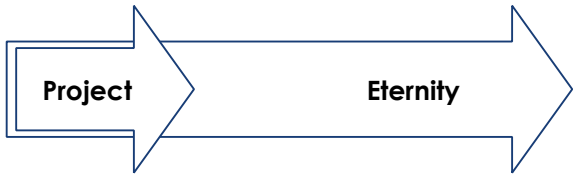
Link to a movie that I'd like you to watch, showing all the things that currently often go wrong when data is re-used.
Data is available on-request, but is actually difficult to find; the software that can read is is no longer available, the data columns are not understandable without a post-doc who has left and can no longer be contacted.

Before explaining what we do, we first visit a Dystopian world, unfortunately not to far from the truth in some cases

https://www.youtube.com/watch?v=N2zK3sAtr-4

Although this is brought as a bad example, there are surely parts you recognize.
I certainly recognize it.

Data Lifetime



So there is more data. Also we want to be able to reuse it for a longer period. It needs data management, or data stewardship.
Data Stewardship vs Data Management naming pitfalls: "data management" is sometimes considered to **end** at project end. "data stewardship" is sometimes considered to **start** at project end.
We call it Data Stewardship, and for us this includes Data Management
For me it is any operation that handles digital data. Some others include data on paper and/or samples

---

DTL Definition of Data Stewardship

Responsible planning and executing of all actions on digital data before, during and after a research project, with the aim of optimizing the usability, reusability and reproducibility of the resulting data

DTL Definition of Data Stewardship

Responsible planning and executing of all actions on digital data before, during and after a research project, with the aim of optimizing the usability, **reusability** and reproducibility of the resulting data

Lets focus on that Re-usability for a moment

Reusable

We want to make data re-usable. In first instance that sounds like an action done for others: the reusers of our data.

However, I draw the word in two colors for a reason: making the data re-usable early enough also makes it usable in the project.

Furthermore, researchers are the first re-users of their own data three times: tomorrow when you want to continue, in three months when a reviewer ask you to try a small change, and in 2 years when your postdoc leaves and a new one needs to pick up the work.

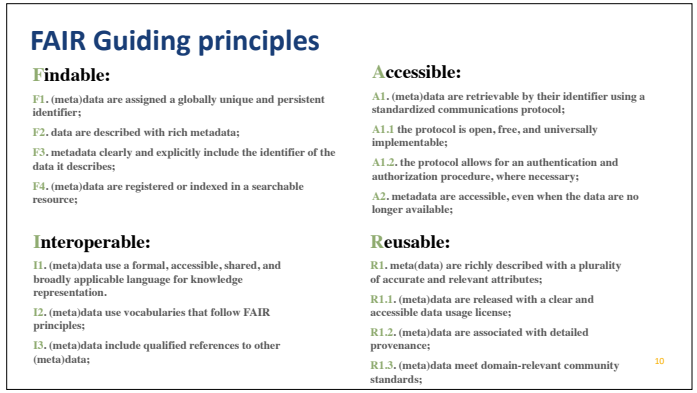
Well described re-usable data saves the day every time.

FAIR

Re-usable happens to be the last of 4 letters of FAIR: the goal that data from research should be made Findable Accessible Interoperable and Reusable for humans AND Machines.

Concept of FAIR data was developed in 2014 in a meeting in Leiden, NL

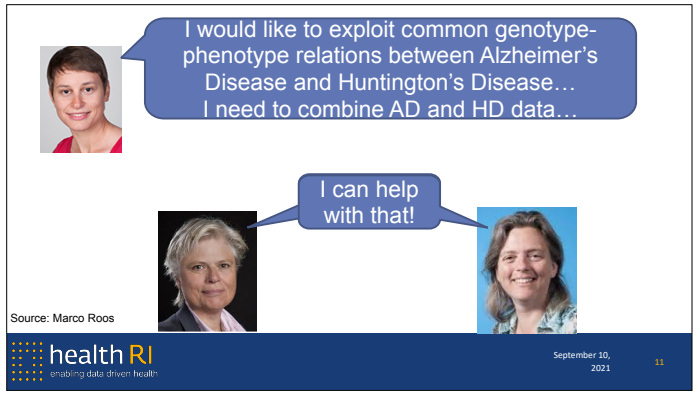
published in 2016, and used by G7 and G20 among others



Findable: if the data is there, but it can't be located, discovered by someone who needs it, it is basically wasted.
Accessible: Protocol to access the data. And important that "metadata" should be kept (as kind of publication) even when the data has disappeared (unless you keep paying it will disappear.
Interoperable: The biggest challenge to make it all work together
Reusable: Metadata on what you are allowed to do with the data, licensing, fees, and provenance.

Principles are very generic, there is no hint how this can be done in practice.
Later in this talk I will give some practical examples what this can mean for research projects

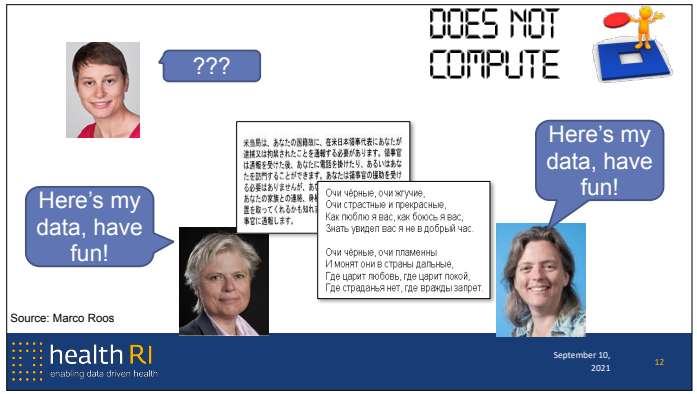
Reference: Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)



What does It mean to be re-usable?

Re-using data often means combining data from different sources. But this is less "automatically possible" than we would like.

In this example from Marco Roos from Leiden, a researcher wants to combine data from different diseases. The data is available.



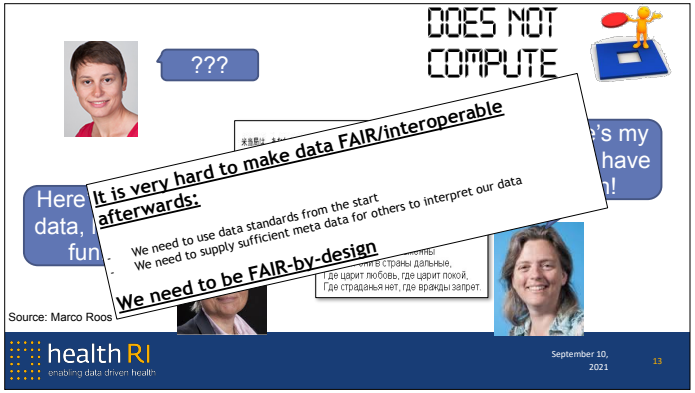
But the data is not easily joined together!

It would have helped to choose the same standard, but this is not 100% necessary: if different languages are used a translation is certainly possible. I'm not only meaning human languages here but any kind of differences in how things can be addressed, like e.g. gender or dates.

What is most important is that all "assumptions" are very well documented. We need metadata. This is the case when a translation is needed but also when the same language is used.

We need to know what the data are, how they were obtained, how they were coded, and maybe even why.

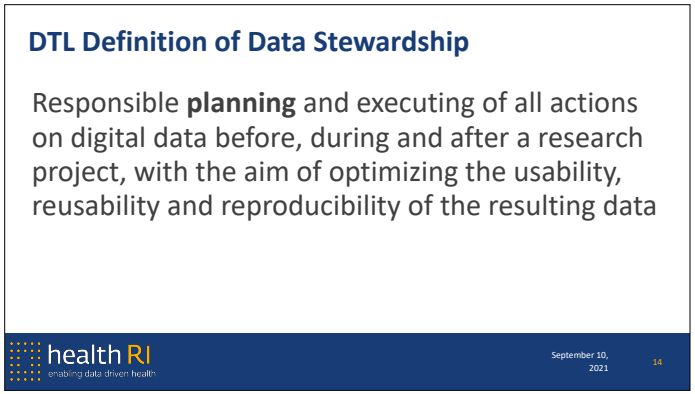
It must be impossible to misinterpret!



This FAIR interoperability is not easy to add later. It is much easier done early. There where the data is collected, the metadata also is known. And remember that that also helps yourself as the first reuser of your own data.

Data has become such an important part of projects that we can no longer afford to just do stuff and correct it if it goes wrong. We actually need to start planning the data stewardship as part of the planning of a project.

.

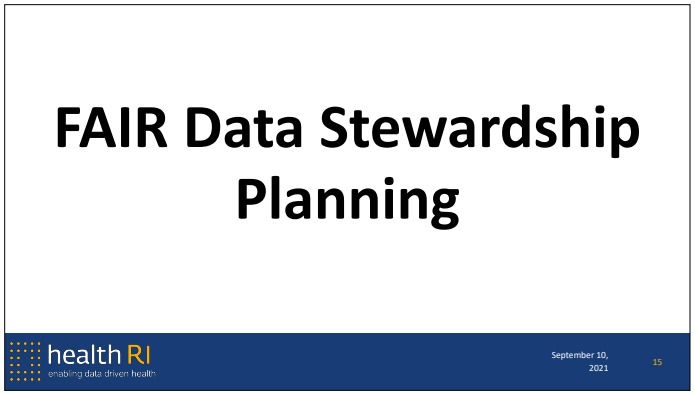


This is the reason for the need for the word ”planning” in the definition of data stewardship.

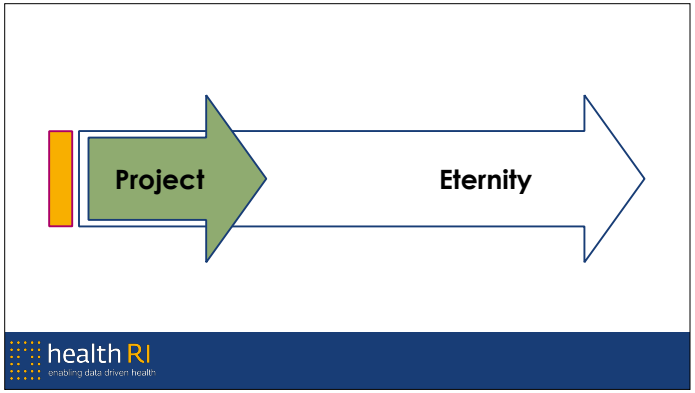
If you plan, there is much less that can go wrong. Chemists plan laboratory safety, and DMP amounts to data safety. Without lab safety we have accidents, many data accidents happen without planning, this is often not yet recognized as avoidable.

In addition, Horizon Europe, more than Horizon 2020, now really wants every project to do data stewardship planning

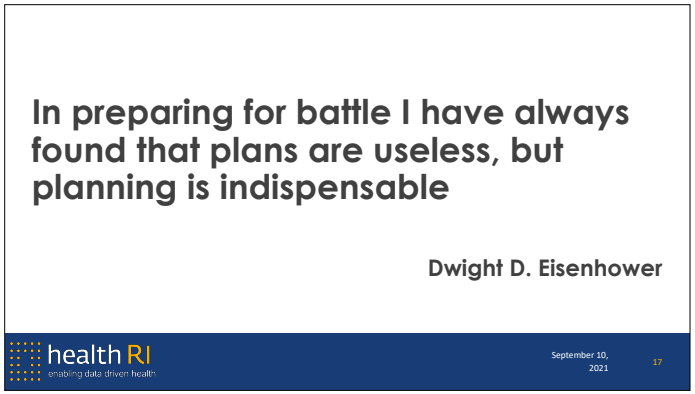
---



Combining the elements, we promote the idea of FAIR Data Stewardship Planning: Using the FAIR principles do decide on Data Stewardship decisions, and starting with that process before the project starts



Many funders want us to make DMPs. The right moment to do that is starting before the project, and updating the plan during the project. We need to embrace reality: the plans change over time. But the DMP should be updated to represent changes in the actual way things are done.

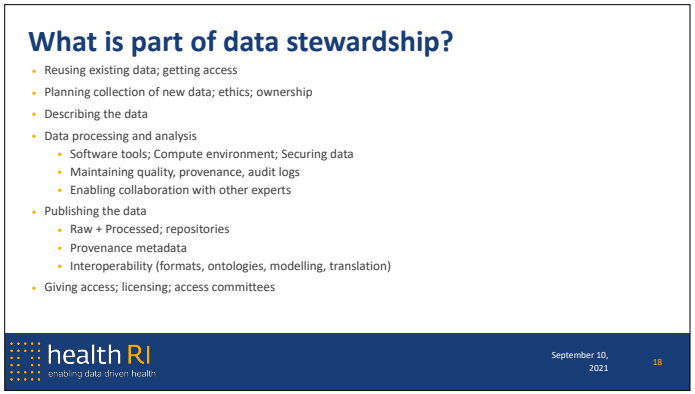


DMP's change over time.

What can we learn from warfare? It is normal that Plans change.

Alternative: "no battle plan survives contact with the enemy" by Helmuth von Moltke
Or: When your plan meets the real world, the real world wins

Remember that no matter how carefully you plan your work, reality will always be different from what you expected.



Data Stewardship has many components, some of which are the expertise of others.

To get the most value out of your data and to run the least data risks, all of these need to be arranged.

All of these are different for each project. All of these benefit from planning. Call upon experts.

I will now pass each of the letters of FAIR and give you some first steps.

Findable

- Choose Repository
  - Domain specific
  - Institutional
  - National
  - Special
- Make sure to get and use persistent identifiers
- Describe what the data **is**, not why you collected it
- Register the data in a catalogue, where others would look for it

* Data needs to come into a repository
* Special = Made for the project. Note Accessibility!
* Different repositories for different data?
* Preferences from funder or institute? Budget affected!
* Go talk to repositories! They may require format, metadata. They can help!

* Persistent identifier like DOI, Help you get credit

* Especially institutional or special: catalog it!
* Keywords for re-use. What is in there, not why you collected
* Keywords for subsets?
* Librarians/archivists can help! Go look for them!
* Talk to the catalog people early to find out what they need.

Accessible

- Assure Longevity
- Check Legal Conditions
- Limit the Embargo
- Ensure proper ICT procedures

- Longevity: 20% decay per year on links to web sites. Certified repositories! Transfer if they stop.
- Software longevity
- Repositories often work with one-time-fee.
- Privacy sensitive data: Not coupled to person! Clear rules! Find a data access committee
- Non-legal reasons: Limited time! not "once we get the patent" or "after the publication"
- ICT procedures: convince reviewer that you won't lose (track of) the data. Professional?

With accessibility think about machines that need to couple a lot of data sources, the protocols need to be really standard to make that easy.

Interoperable

- Format
- Terminology
- Sufficient metadata

Brussels ← → Bruxelles
Cancer ← → Malignant Neoplasm

* Format
* Terms: (controlled) vocabulary. Or an (international) coding system. With relations: Ontology.
* Examples after click
* Use what others use, or map
* All fields: even if you do genetics, store sample locations like climate scientists.
* Simple fields like date and sex too!

There will be lots of development, demands on interoperability and benefits will grow!

•1854 London Cholera outbreak, John Snow made a map and identified a water pump that was the origin of the infection
•Arabidopsis study where correlation was made with height through google maps API.
•Sufficient metadata to interpret the data unambiguously.

Reusable

- Document provenance
- Use *minimal metadata* standards
- Choose a liberal license
- Make sure results are not only narrated

- Annotate to help self and others, avoid clarification requests
- How was data obtained, provenance. Instrument settings. Automate collection! Helps yourself!
- Minimal metadata. Quite extensive, and volunteer to do optional too. Field specific+DC.
- License. Are you obliging people to Cite? Forbid commercial use? Such clauses have Consequences. Unnecessary restrictions!
- See the example of the Gauss curve. Which nobody cites any more!
- Ewan Birney insight: other scientists will cite, they need to prove they got data from a reliable source!

Do not encode knowledge only in narrative. Please help to make text mining obsolete.

* use experts, be experts.

Thought

Making the actual DSP requires that you think about it. You can't simply copy a DMP from somewhere.
How do you manage to make your data FAIR
Some aspects can be more difficult than you imagined. Especially where it is far from your own specialty.

You may have a well maintained photo library at home. One that you are proud of.
But: Maintaining data in the lab is more complicated than a photo library. (1) It is larger (2) It is varied (3) It is operated on by different people
And even in a photo library you are sometimes looking forever.

Regarding this photo library: everybody makes the same mistakes.
You should really make use of the expertise of others. Look for the experts!

Irritant
Painful
Dangerous

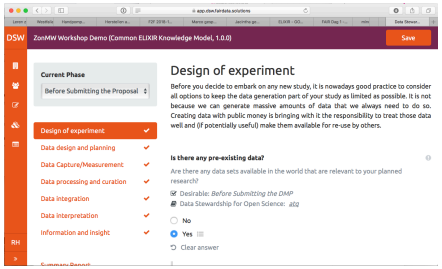
Broad expertise is needed: from (library, IT) and from field experts
Not all the data related expertise may be in your existing network

Service providers make a list. Does that work? No, Lists only solve the easiest problem: finding back something you already know that exists.

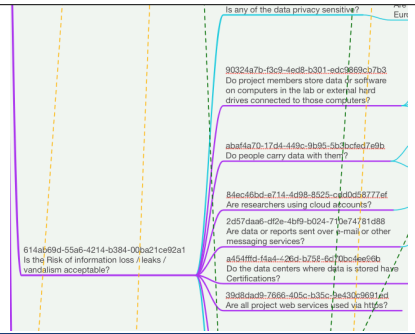
It is irritant if you don't know where to find the expertise. Lists solve that.
It is painful if you don't know it exists. Lists don't solve that.
If is dangerous if you don't know you need it. Lists don't solve that.

Case in point: How do you get your apps on your phone? You ask a friend, an expert.
I think I can help in another way.
With a tool that we are developing in ELIXIR. A tool that behaves like an expert.

Data Stewardship Wizard

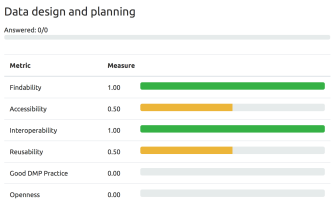


In ELIXIR-NL we work together with ELIXIR-CZ on a tool that can guide researchers to good data management: the wizard. The goal of the wizard is the PLANNING, not the PLAN (which is useless, you remember?)
It attempts to ask the first few questions a broad set of experts would ask and point to more information.
It opens relevant questions (only)
It contains more detailed guidance on some topics.
It contains links to Barend's book pages
It measures the FAIRness of the result per chapter.



The questions in the questionnaire are based on a mind map I had first been collecting over 5 years in interviews with experts.

Includes metrics



For each answer in each question, the system also encodes Six objectives.
Not as an exam, but with guidance.

Next to this system there are other systems out there to calculate FAIRness, often once data sets exist somewhere out there.

FAIR Data Maturity Model

- For the EC, the Research Data Alliance built a "FAIR Data Maturity Model"
  - https://doi.org/10.15497/RDA00050
- Series of yes/no checks you can do on your data
- Some tools exist to do this automatically (subsets of the full model)

- It is not important to measure how FAIR compliant your are, but to identify how you could easily become more FAIR in the future

A model like this is very good to make very clear what the original principles mean. But it is useless to do automatic and blind checking to get a "score", because they are incomparable! Instead, it is better to check where you can make changes to your data that are easy to do and valuable for FAIRness. An investment analysis.

Data Stewardship Planning

✅ F

✅ A

✅ I

✅ R

Proper Data Stewardship makes and keeps data FAIR, during and after the project.

- (Researchers do not want to make DSP? Do they want to make Papers? Both belong to science!)
- (Requirements on FAIR will grow, FAIR is a good start)
- FAIR approach to DMP will not only benefit others, but also yourself.
- At the end of the project, the DSP has become a map of the road you walked.

Perspectives from a DMP reviewer

Reviewing many DMPs, a reviewer very often encounters planning that is badly described, could never work in practice, or could go really badly if executed as described. One can not help getting a bit cynical reading "intentions".

Reviewer perspective 1: “F” example

DMP: “The data can be located through the DOI”

Reviewer: Hopefully not ”the” DOI of “the” paper?
Reviewer: How about availability of the data if ”the” paper is not accepted?

Reviewer perspective 2: ”A” example

DMP: “Data are stored centrally in our institute, with appropriate backup facilities”

Reviewer: Do we know what the backup protects against? Mostly it is equipment failure, not human error!
Reviewer remembers a project where 800 TB backup needed to be restored, delaying a project by 3 weeks more than needed.
Reviewer: How often are backups made? How long to restore? Is that acceptable to the project? Was this given thought, or purely assumed?

Reviewer perspective 3: ”A” example

DMP: “Data can be obtained by sending a research proposal to our PI”

Reviewer: What if the PI moves on?
Reviewer: Is the PI judging those proposals on gut feeling only?
Reviewer: How does a machine know what the data can be used for (consent)?

Reviewer perspective 4: ”I” example

DMP: “We do not collect metadata”

Reviewer: Everybody has and needs metadata! Who is collecting the data? What kind of instrument is being used? What age was the subject? What disease did you study? Was this a disease subject or a control?

The temperature is 25”? Where? What unit? How measured? How accurate? When?
Given: distinction with data is often weak. In the eye of the user.

Reviewer perspective 5: ”I” example

DMP: “Data are available in XLS (or TXT or SPSS). Images in JPG (or TIFF)”

Reviewer: An XLS is not self-documenting. What column is what?
Reviewer: Where is the image metadata?

Reviewer perspective 6: ”R” example

DMP: “Data are freely available for non-commercial use”

Reviewer: Does the researcher actually own the copyright?
Reviewer: Can non-commercial be enforced? What would have happened if Gauss would have made his research available for non-commercial use?
Reviewer: Is this creating the best societal value?

Reviewer perspective 7: ”R” example

DMP: “Researchers will use a lab-journal to note instrument settings”

Reviewer: How will re-users of the data have access to this meta-data?
Reviewer: Will this be machine readable?

A good DMP helps!

This is not only cynicism.

I sincerely believe that a DMP can be helpful for research. Thinking beforehand can avoid expensive mistakes, and timely discussions with experts can reveal options to do things more efficiently. Data being truly as reusable as possible for others later also helps making the data more re-usable for yourself.