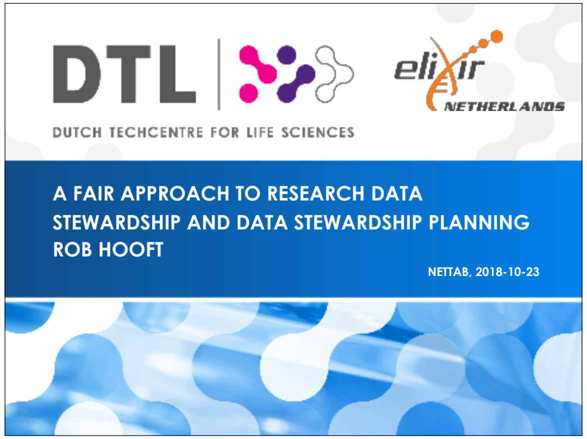


DTL

DUTCH TECHCENTRE FOR LIFE SCIENCES

elixir NETHERLANDS

A FAIR APPROACH TO RESEARCH DATA STEWARDSHIP AND DATA STEWARDSHIP PLANNING
ROB HOOFT

NETTAB, 2018-10-23

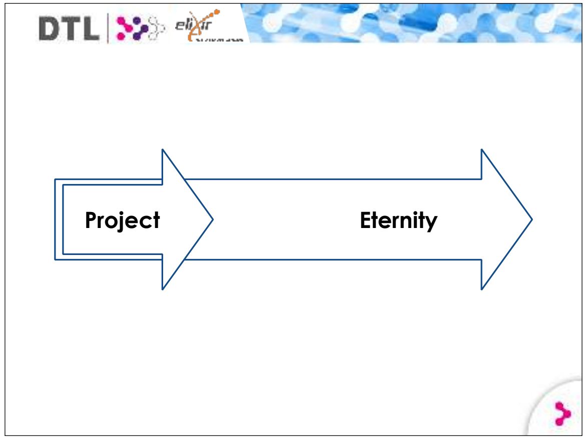
Dystopia

But your work is in PubMed Central and was funded by NIH. That is true!

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.



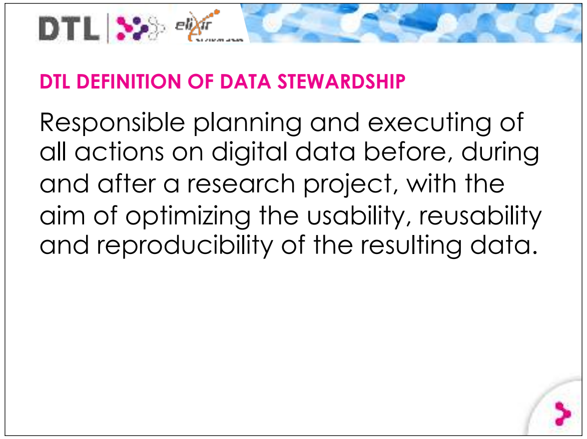
Project

Eternity

We want to be able to reuse data 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.

Data Stewardship requires good Data Management.

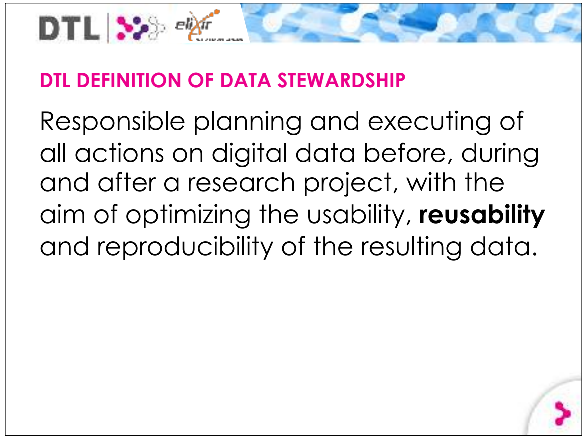


DTL DEFINITION OF DATA STEWARDSHIP

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.

We call it Data Stewardship, and for us this includes Data Management

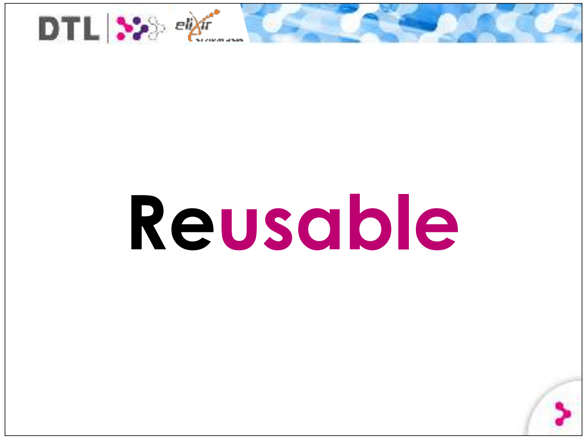
For me it is any operation that handles digital data. Some others include data on paper and/or samples.



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.

First lets have a closer look at the word reusability.



Reusable

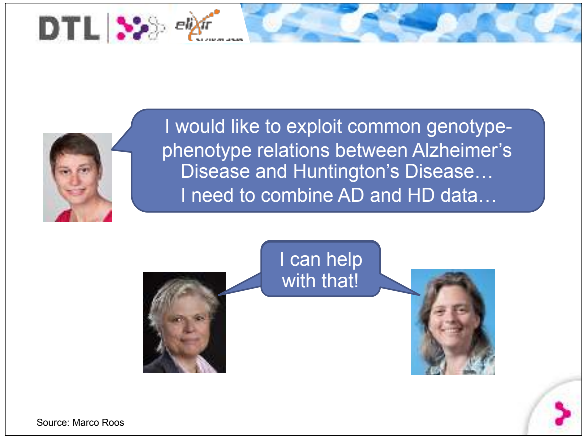
Our project work should aim to make the data Reusable.

- Reusable for others
- Usable and Reusable for ourselves
  - Now usable: data is complete and well described. You don't need to search for all the metadata at the end, but properly maintain it.
  - Later reusable: because you won't remember when the reviewer asks, and because you struggle to reproduce
  - Even later reusable when your postdoc is replaced by a successor

FAIR

To make our data re-usable, we need to make it Findable, Accessible, Interoperable.
It is like a four step ladder.

And thinking about this really helps also making decisions in a project.

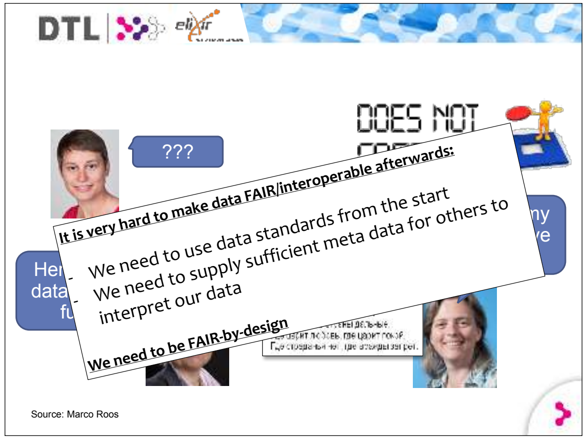


I would like to exploit common genotype-phenotype relations between Alzheimer's Disease and Huntington's Disease…
I need to combine AD and HD data…
I can help with that!

Source: Marco Roos

What does It mean to be re-usable?
Interoperability is not something that can be added as an afterthought. It should be planned for, that really saves time.

Data management is the answer to a complicated question, and is not easy! Needs thought.



???
DOES NOT
Her
data
fu
ny
ve

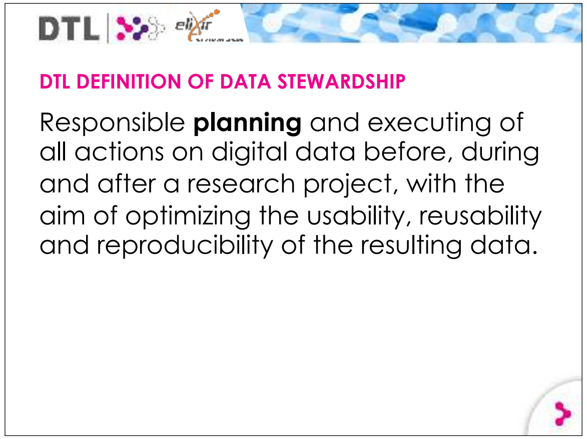
It is very hard to make data FAIR/interoperable afterwards:

- We need to use data standards from the start
- We need to supply sufficient meta data for others to interpret our data

We need to be FAIR-by-design

Source: Marco Roos

Data interoperability is more than technical standards: extensive meta data are needed

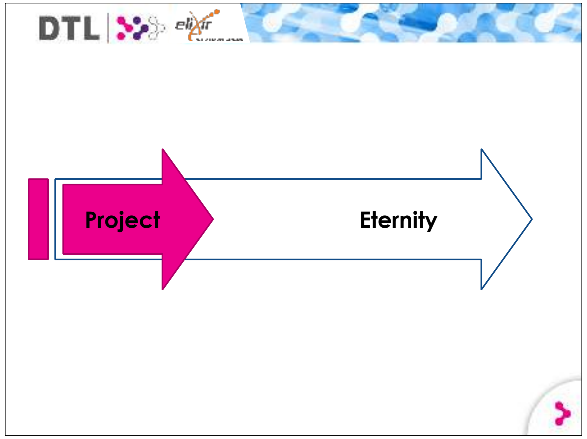


DTL DEFINITION OF DATA STEWARDSHIP

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.

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

FAIR Data Stewardship Planning

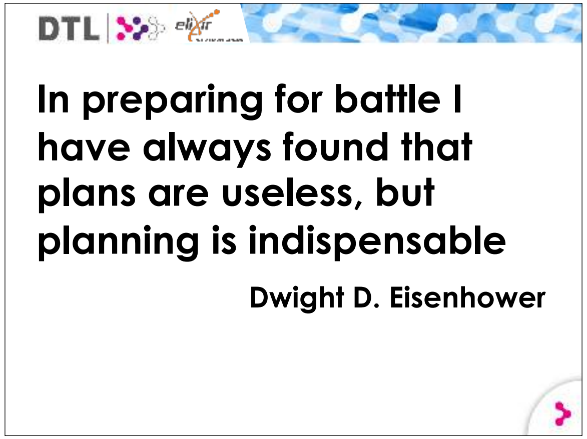


Lets first talk about the timing of the DMP activity.

Different funders have different ideas about when a DSP should be made. Some do it before the project starts, but that is hard because that activity is unfunded; it discounts how much work is actually going into good planning. Some ask it as a deliverable, but that makes the budget really hard in some cases. The best way is to start planning before the project proposal, but continue the DMP in the project.

To be able to deal with changes in the project, the data stewardship plan should be updated all along with the project. That way it will also function as part of the documentation of the data after the project is finished.

Big project? Hire a data expert, one of the 500k the EOSC calls for. A data steward should know a lot, and also know where to find the experts in your institute that can help further.



In preparing for battle I have always found that plans are useless, but planning is indispensable

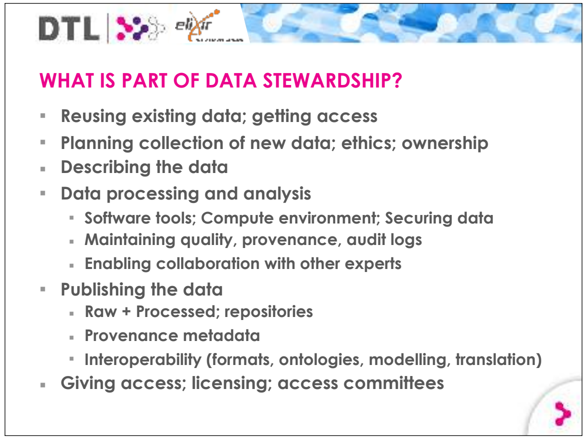
Dwight D. Eisenhower

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.



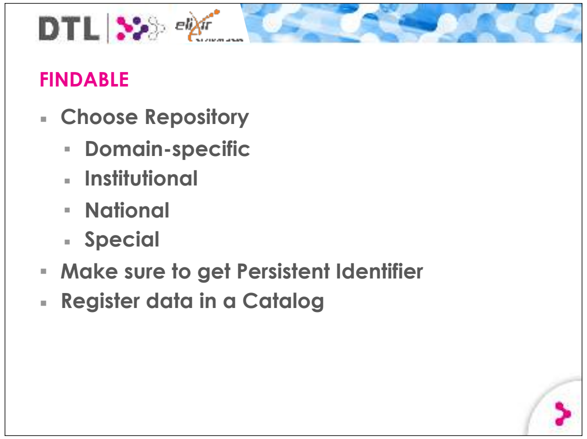
WHAT IS PART OF DATA STEWARDSHIP?

- Reusing existing data; getting access
- Planning collection of new data; ethics; ownership
- Describing the data
- Data processing and analysis
  - Software tools; Compute environment; Securing data
  - Maintaining quality, provenance, audit logs
  - Enabling collaboration with other experts
- Publishing the data
  - Raw + Processed; repositories
  - Provenance metadata
  - Interoperability (formats, ontologies, modelling, translation)
- Giving access; licensing; access committees

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 very quickly pass examples.



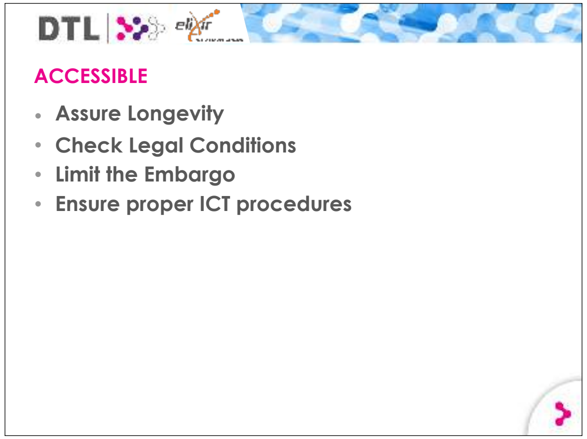
FINDABLE

- Choose Repository
  - Domain-specific
  - Institutional
  - National
  - Special
- Make sure to get Persistent Identifier
- Register data in a Catalog

* 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

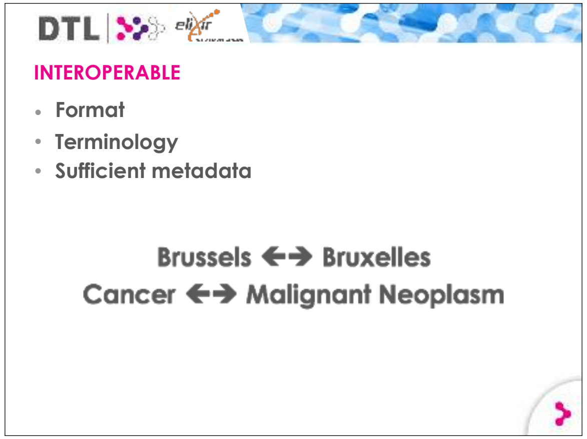
* Especially institutional or special: catalog it!
* Keywords for re-use. What is in there, not why you collected
* Keywords for subsets?
* Librarians/archivists can help! Go look for them!
* Talk to the catalog people early to find out what they need.



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?



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

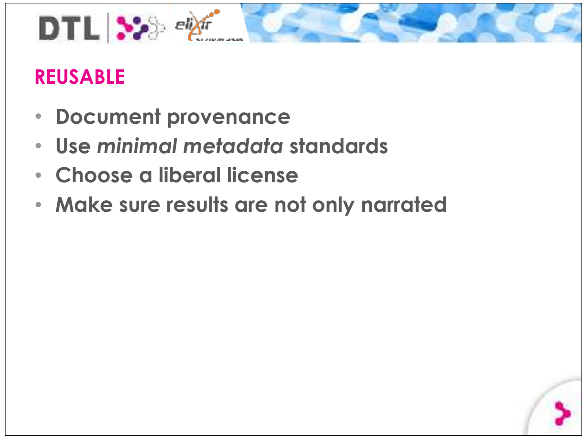
* All fields: even if you do genetics, store sample locations like climate scientists.

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

Example

•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.

•You need sufficient metadata to be able to interpret the data



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.

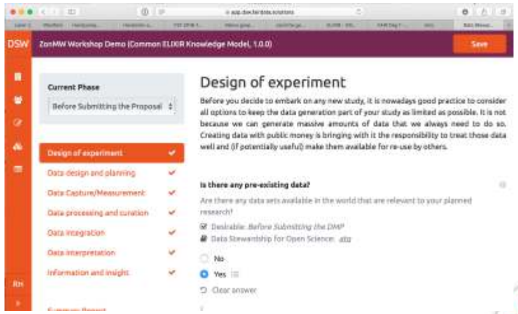
If is dangerous if you don't know you need it. Lists don't solve that.

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


DSW
ZonMW Workshop Demo (Common ELIXIR Knowledge Model, 1.0.0)
Save
Current Phase
Before Submitting the Proposal
Design of experiment ✓
Data design and planning ✓
Data Capture/Measurement ✓
Data processing and curation ✓
Data integration ✓
Data interpretation ✓
Information and insight ✓
RH

Design of experiment

Before you decide to embark on any new study, it is nowadays good practice to consider all options to keep the data generation part of your study as limited as possible. It is not because we can generate massive amounts of data that we always need to do so. Creating data with public money is bringing with it the responsibility to treat those data well and (if potentially useful) make them available for re-use by others.

Is there any pre-existing data?

Are there any data sets available in the world that are relevant to your planned research?

Desirable: *Before Submitting the DMP*

Data Stewardship for Open Science: ato

○ No

● Yes

Clear answer

In ELIXIR-NL we work together with ELIXIR-CZ on a tool that can guide researchers to good data management: the wizard. https://dsw.fairdata.solutions/

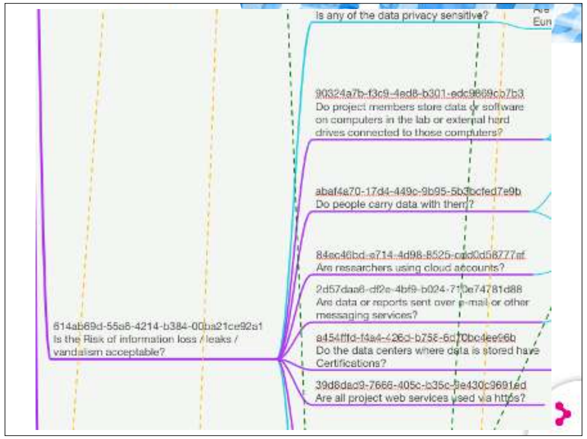
The wizard opens questions that are relevant. It attempts to ask the first few questions a broad set of experts would ask and point to more information.

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.



Is any of the data privacy sensitive?

Are Eur

90324a7b-f3c9-4ed8-b301-edc9869cb7b3
Do project members store data or software on computers in the lab or external hard drives connected to those computers?

abaf4a70-17d4-449c-9b95-5b3bcfed7e9b
Do people carry data with them?

84ec46bd-e714-4d98-8525-cad0d58777af
Are researchers using cloud accounts?

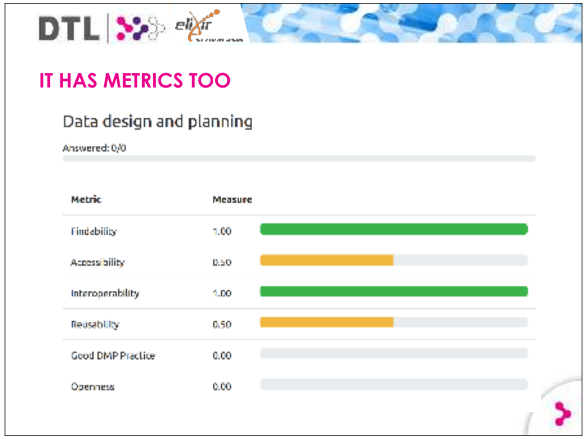
2d57daa6-df2e-4bf9-b024-710e74781d88
Are data or reports sent over e-mail or other messaging services?

a454fffd-f4a4-426d-b758-6d70bc4ee96b
Do the data centers where data is stored have Certifications?

39d8dad9-7666-405c-b35c-9e430c9691ed
Are all project web services used via https?

614ab69d-55a8-4214-b384-00ba21ce92a1
Is the Risk of information loss / leaks / vandalism acceptable?

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



IT HAS METRICS TOO

Data design and planning

Answered: 0/0

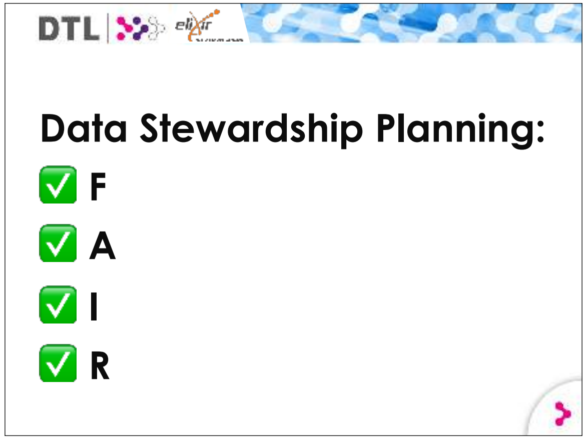
Metric	Measure	
Findability	1.00	🟩🟩🟩🟩🟩🟩🟩🟩🟩🟩
Accessibility	0.50	🟨🟨🟨🟨🟨
Interoperability	1.00	🟩🟩🟩🟩🟩🟩🟩🟩🟩🟩
Reusability	0.50	🟨🟨🟨🟨🟨
Good DMP Practice	0.00	
Openness	0.00	

For each answer in each question, the system also encodes Six objectives.

COMMUNITY AND TRAINING ACTIVITIES

- DTL/ELIXIR-NL:
  - FAIR trainings (for researchers and data stewards)
  - ”Bring Your Own Data” meetings (I/R BYOD)
  - FAIR Hackathons (F/A BYOD)
  - Data Stewards Interest Group (increasingly international)
- ELIXIR DMP Implementation study
  - Inventory of available materials
  - Trainings
- Establishing the Data Stewards job description

See also https://www.dtls.nl/



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.