How the use of controlled vocabularies could be improved to achieve open-science outcomes.

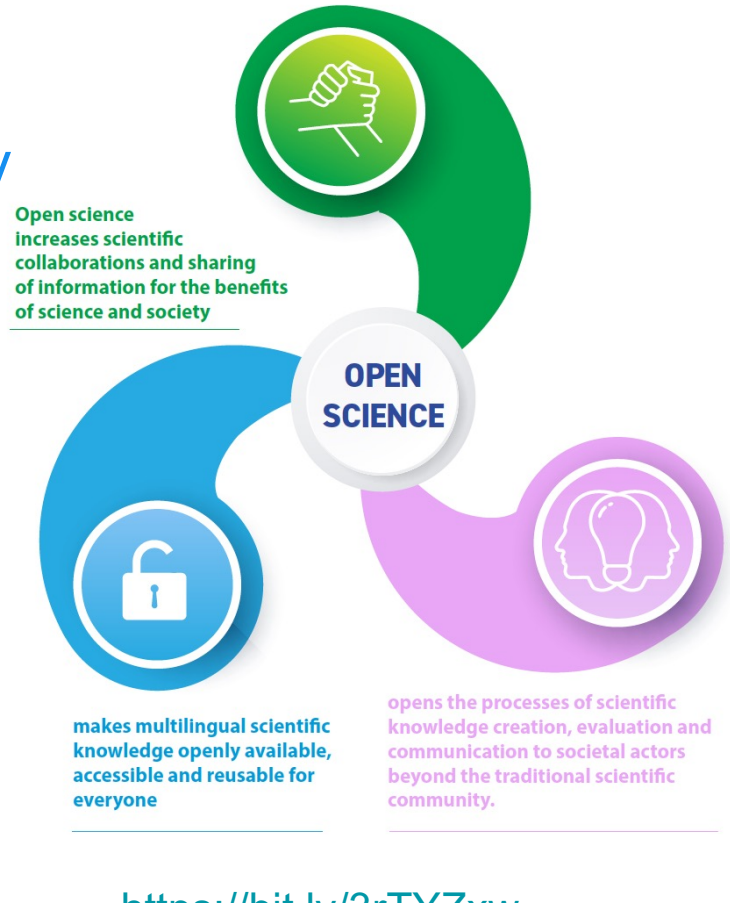
Alison Specht[1], Shelley Stall[2], Yasuhiro Murayama[3], Romain David[4], Margaret O'Brien[5] and the PARSEC Consortium

[1]TERN, University of Queensland, Australia, [2]American Geophysical Union, [3]National Institute of Information and Communications Technology, Japan, [4]European Research Infrastructure on Highly Pathogenic Agents, [5]EDI, University of California, Santa Barbara, USA, www.parsecproject.org.

UNESCO Open Science Recommendations (2021)

Aims:

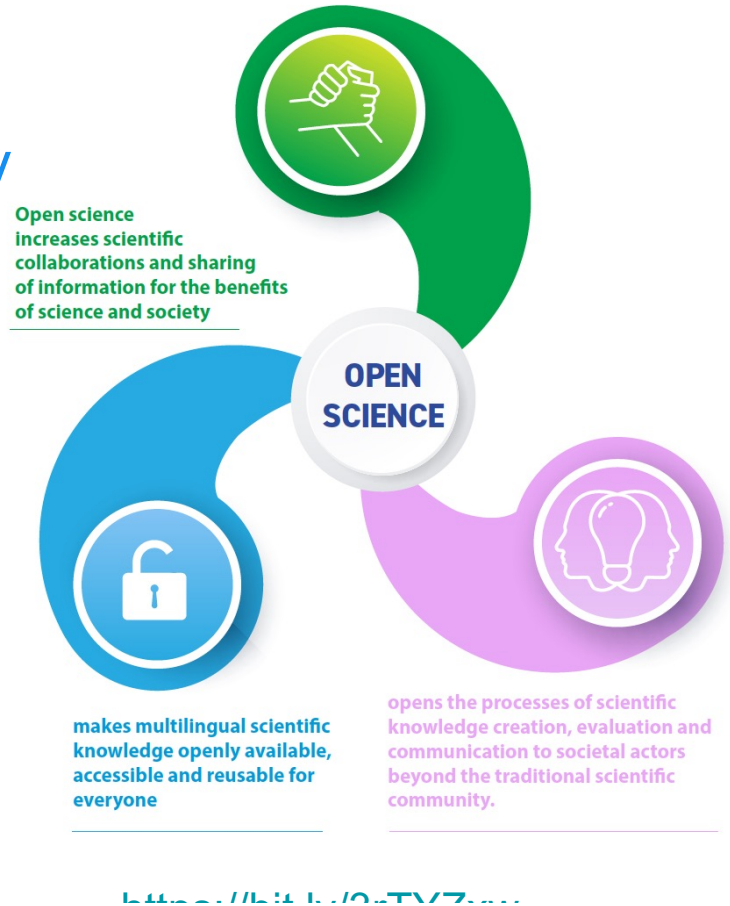
- To make multilingual scientific knowledge openly available, accessible and reusable for everyone,



UNESCO Open Science Recommendations (2021)

Aims:

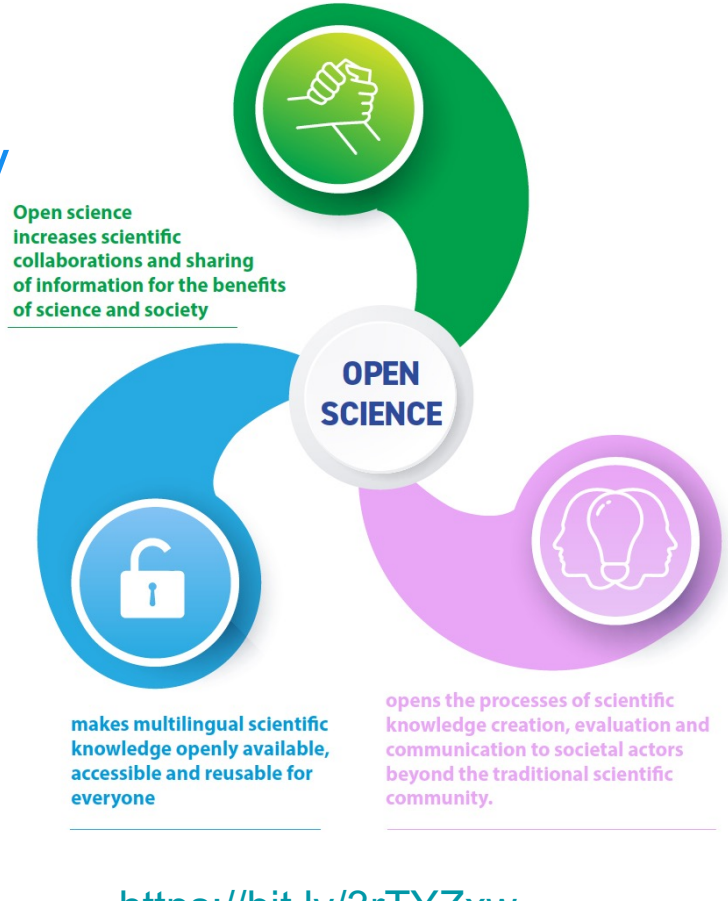
- To make multilingual scientific knowledge openly available, accessible and reusable for everyone,
- To increase scientific collaborations and sharing of information for the benefits of science and society,



UNESCO Open Science Recommendations (2021)

Aims:

- To make multilingual scientific knowledge openly available, accessible and reusable for everyone,
- To increase scientific collaborations and sharing of information for the benefits of science and society, and
- To open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community.



Fundamental to the achievement of open science is sharing the data on which you base your work for others to use.
For this you need to have your data, at the very least, understandable by others.

Where do vocabularies fit in to Open Science?

(thanks to M-A Laporte et al. (2021) Zenodo doi: 10.5281/zenodo.5594693)

Wilkinson, et al. (2016). Sci Data doi: /10.1038/sdata.2016.18

Where do vocabularies fit in to Open Science?

(thanks to M-A Laporte et al. (2021) Zenodo doi: 10.5281/zenodo.5594693)

Notably in Interoperability
(I1, I2, I3)

Wilkinson, et al. (2016). Sci Data doi: /10.1038/sdata.2016.18

Where do vocabularies fit in to Open Science?

(thanks to M-A Laporte et al. (2021) Zenodo doi: 10.5281/zenodo.5594693)

Notably in Interoperability

(I1, I2, I3)

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

Wilkinson, et al. (2016). Sci Data doi: /10.1038/sdata.2016.18

Where do vocabularies fit in to Open Science?

(thanks to M-A Laporte et al. (2021) Zenodo doi: 10.5281/zenodo.5594693)

Notably in Interoperability
(I1, I2, I3)

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2. (meta)data use vocabularies that follow FAIR principles

Wilkinson, et al. (2016). Sci Data doi: /10.1038/sdata.2016.18

Where do vocabularies fit in to Open Science?

(thanks to M-A Laporte et al. (2021) Zenodo doi: 10.5281/zenodo.5594693)

Notably in Interoperability
(I1, I2, I3)

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data.

Wilkinson, et al. (2016). Sci Data doi: /10.1038/sdata.2016.18

Where do vocabularies fit in to Open Science?

(thanks to M-A Laporte et al. (2021) Zenodo doi: 10.5281/zenodo.5594693)

There are different types of vocabularies valued by the community:

- From “weaker” to “stronger” semantics : Glossaries, dictionaries, taxonomies, thesauri, ontologies
- Which type of vocabulary you choose depends on your goal

Ideally, vocabularies supporting FAIR Principles should:

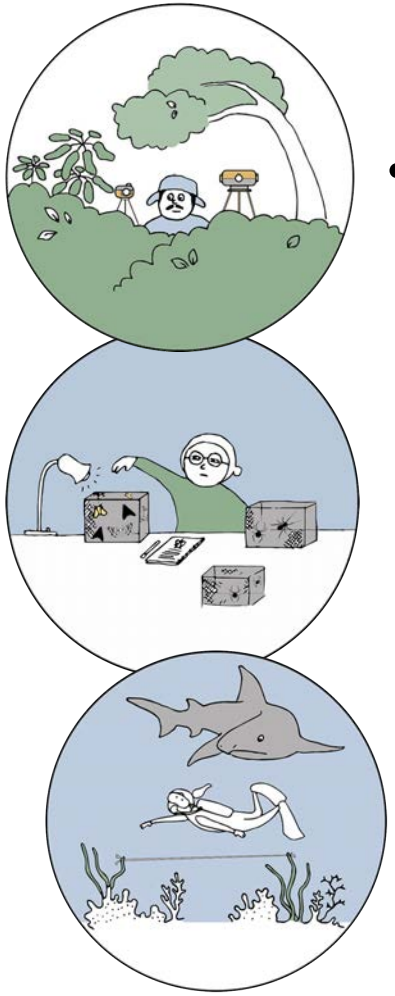
- Provide a shared vocabulary for a domain
- Provide textual definitions
- Standard identifiers (unique, persistent, resolvable by machine)
- Machine Readable format

The challenge (specifically for environmental scientists) is:

Do they appreciate the reason for using standardised vocabularies in achieving open science?

Are there ways to improve acceptance and hence practice?

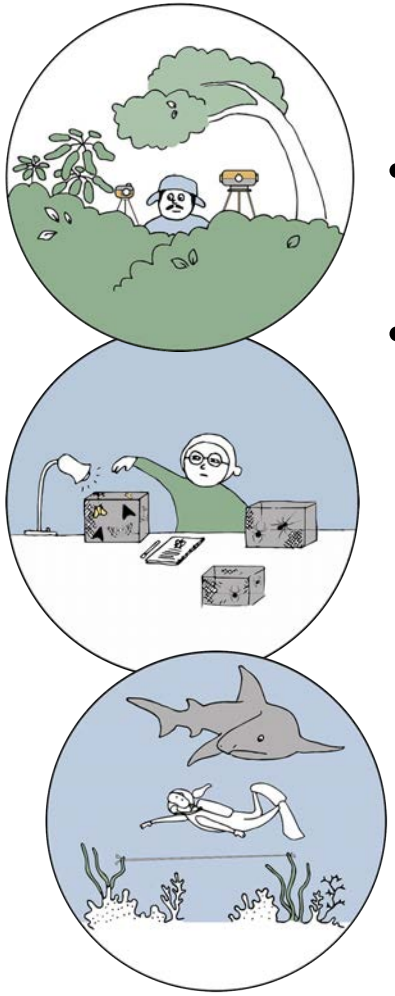
Scene-setting (an example of ecologists)



They

- collect their own data (low reliance on machines) and are really good at R

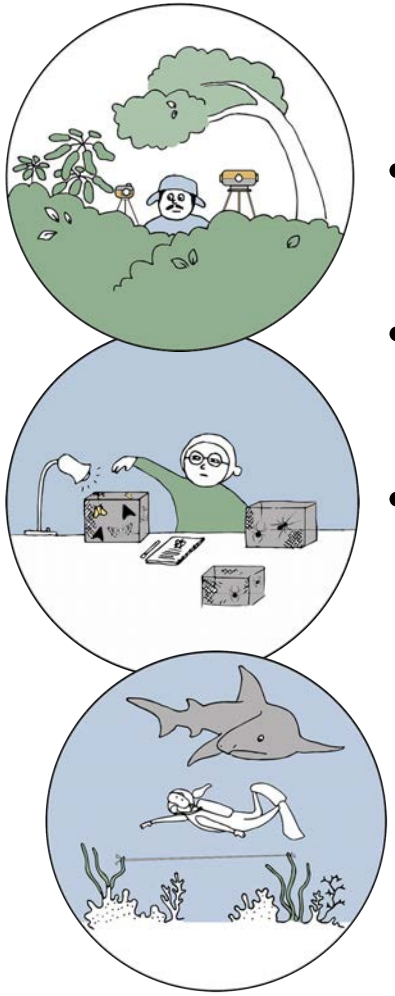
an example of ecologists



They

- collect their own data (low reliance on machines) and are really good at R
- have a strong belief in the value of their work

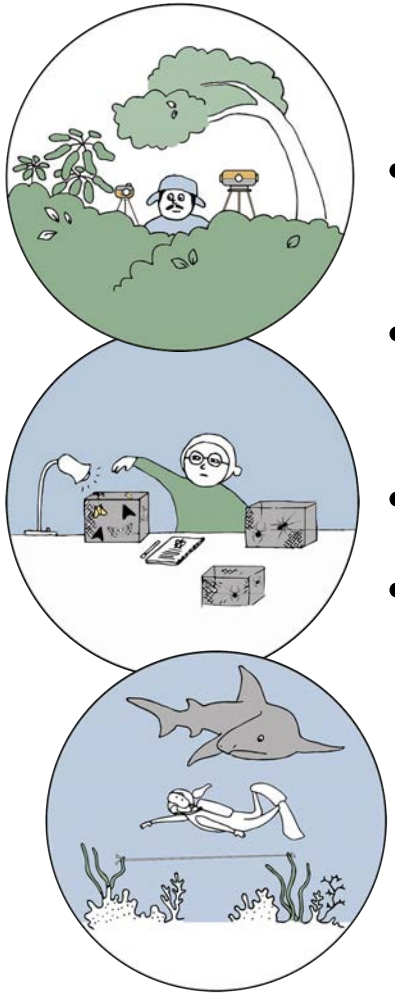
an example of ecologists



They

- collect their own data (low reliance on machines) and are really good at R
- have a strong belief in the value of their work
- work for the common good (low pay)

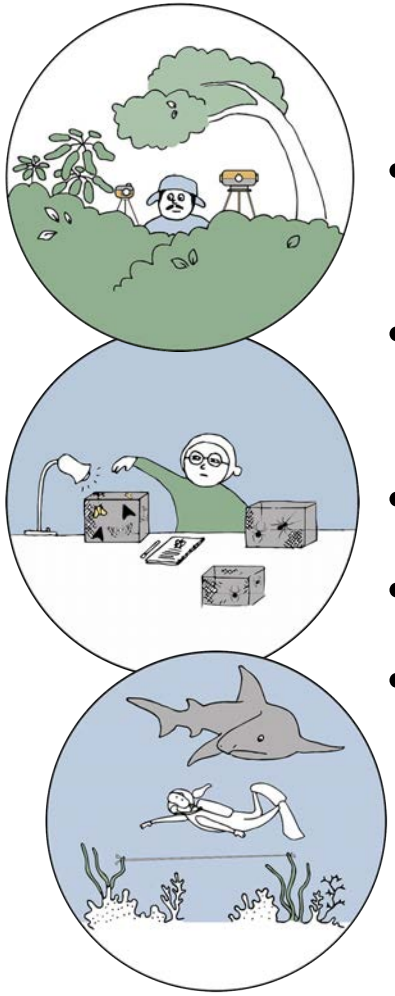
an example of ecologists



They

- collect their own data (low reliance on machines) and are really good at R
- have a strong belief in the value of their work
- work for the common good (low pay)
- are highly variable in their practice

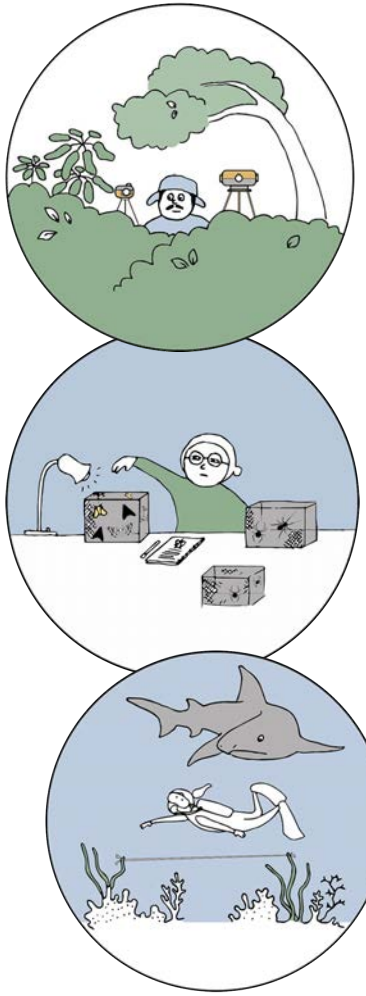
an example of ecologists



They

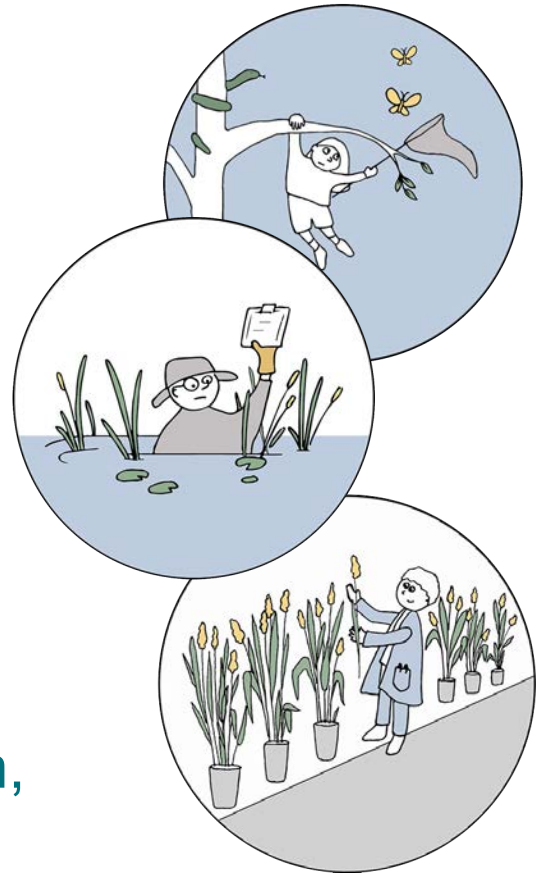
- collect their own data (low reliance on machines) and are really good at R
- have a strong belief in the value of their work
- work for the common good (low pay)
- are highly variable in their practice
- understand the value of long-term data preservation, but

an example of ecologists



They

- collect their own data (low reliance on machines) and are really good at R
- have a strong belief in the value of their work
- work for the common good (low pay)
- are highly variable in their practice
- understand the value of long-term data preservation, but
- tend not to prioritise data standardisation, archiving etc.

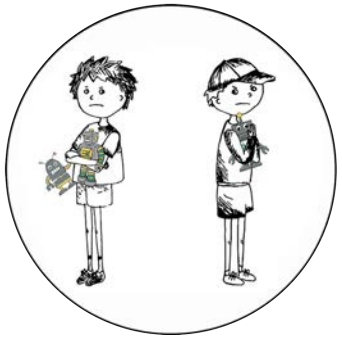


Are these approaches that might work?

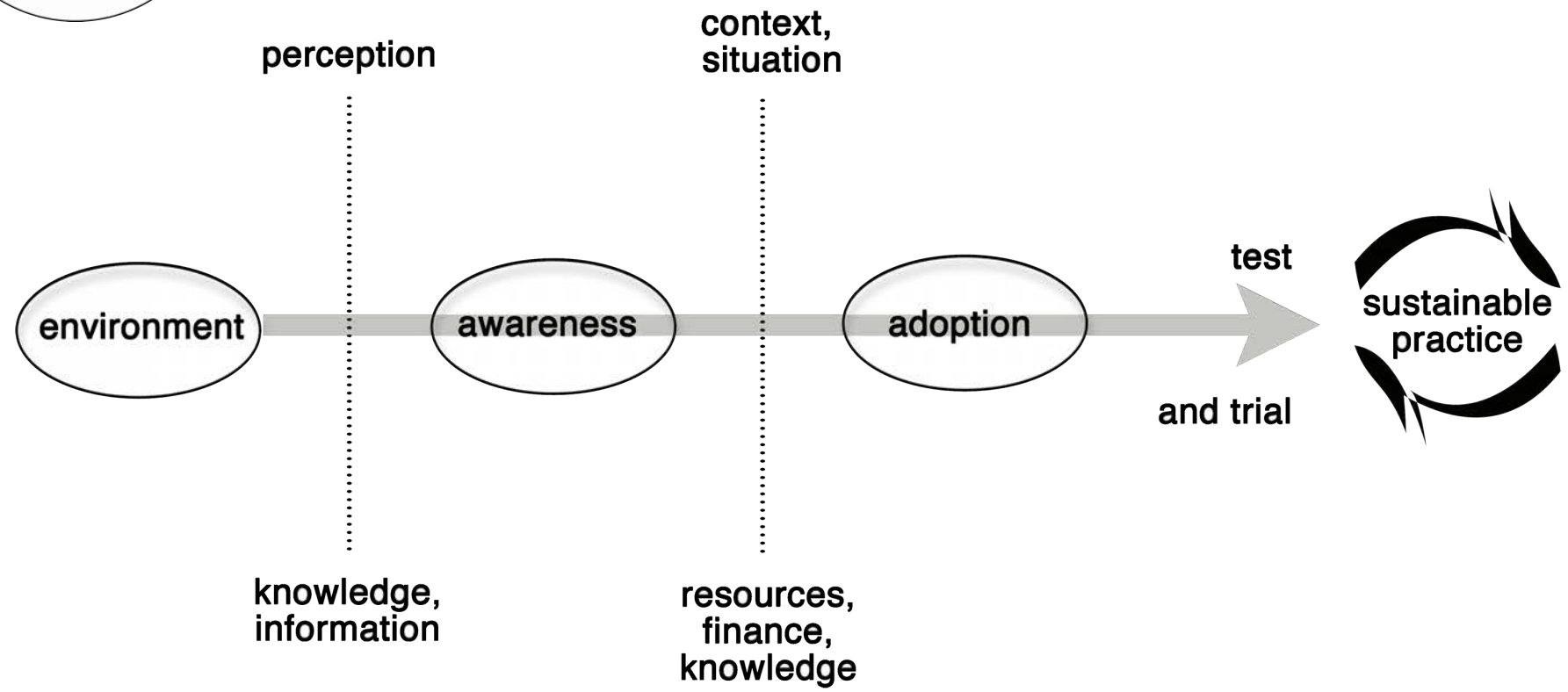
- Provide exemplars?
- Align early with an intended repository (TRUSTed of course)?
- Provide educational packages?
- Are there confounding factors?

Are these approaches that might work?

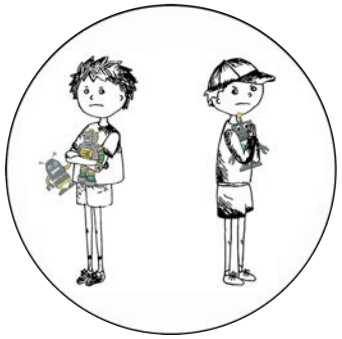
- Provide exemplars?



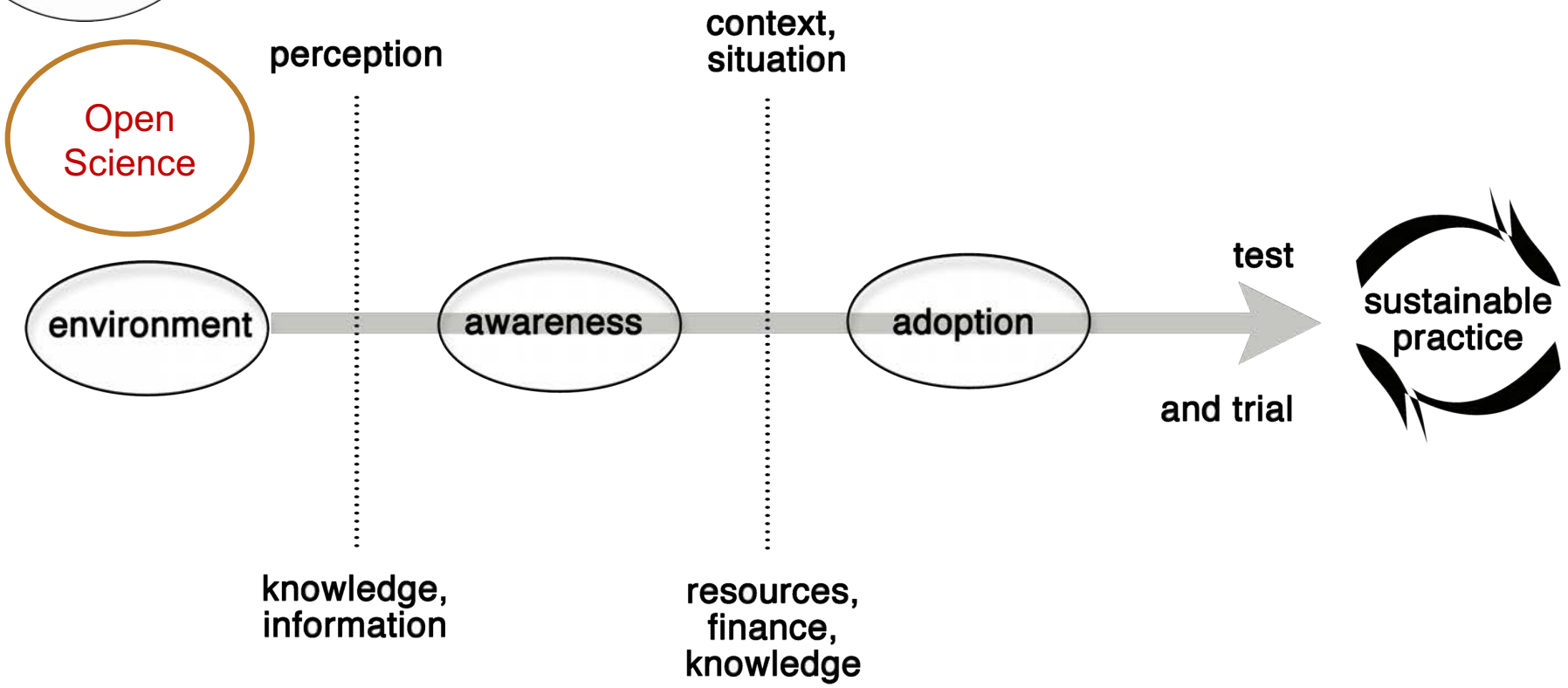
Changing culture and practice



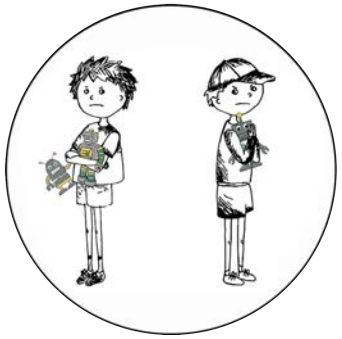
After F. Vanclay (1992)



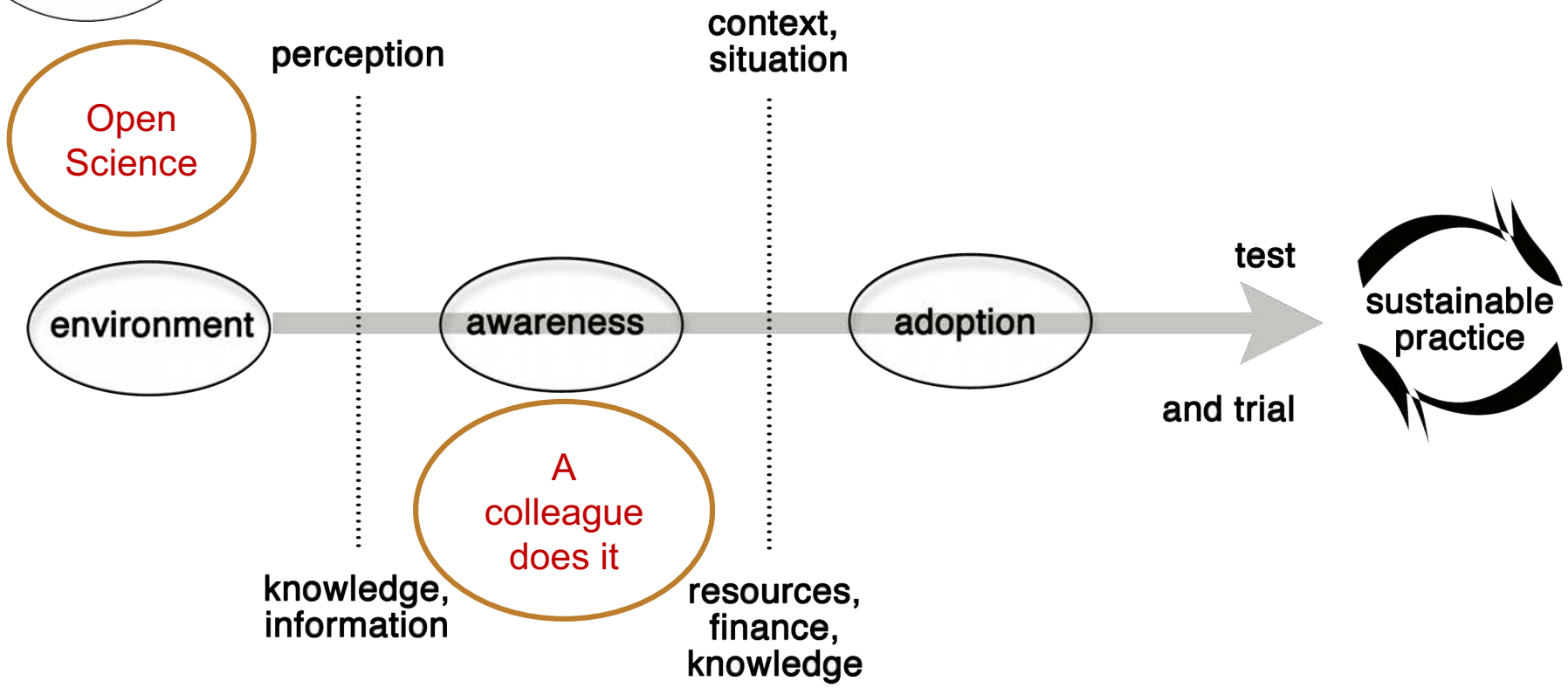
Changing culture and practice



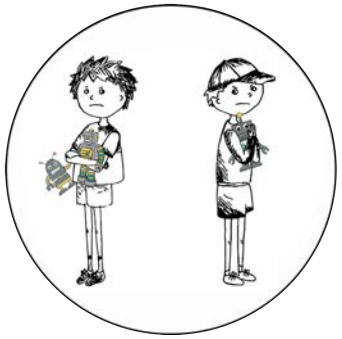
After F. Vanclay (1992)



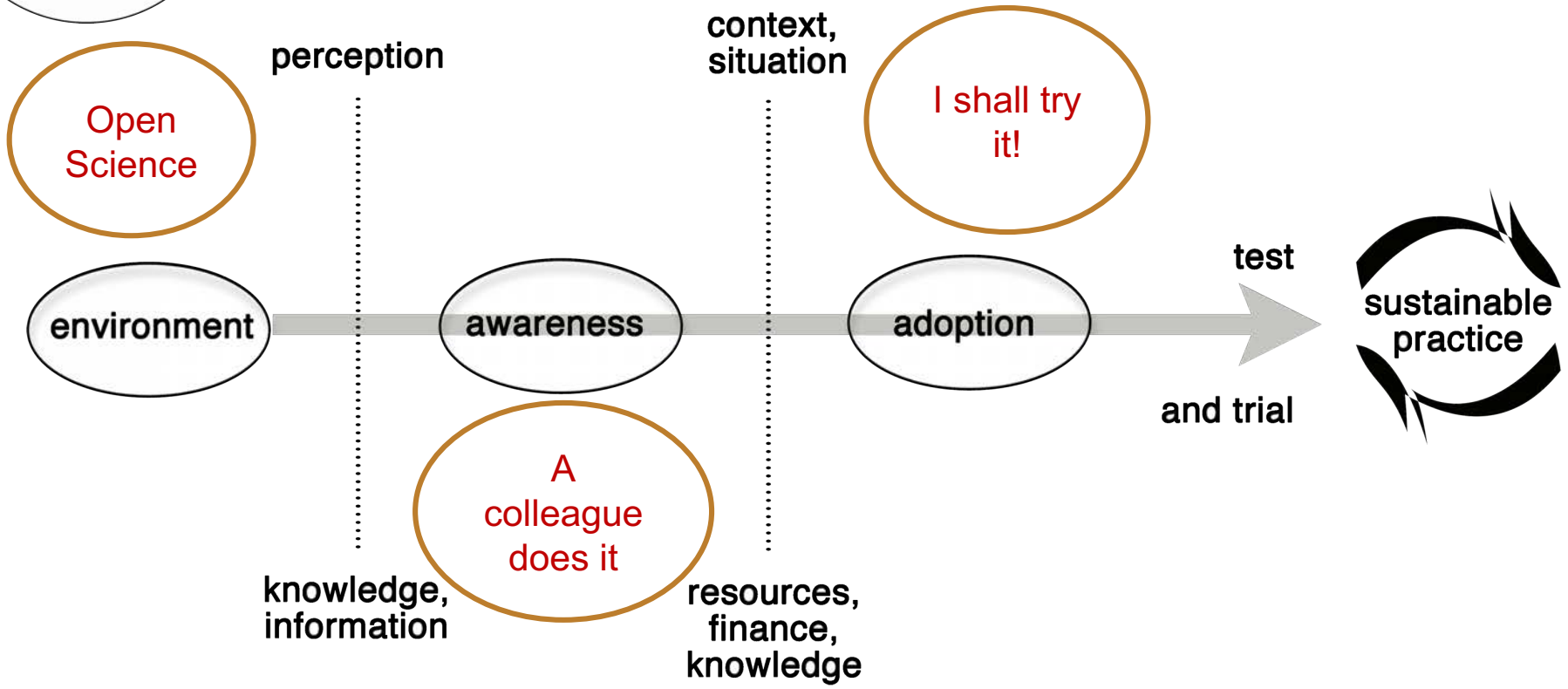
Changing culture and practice



After F. Vanclay (1992)

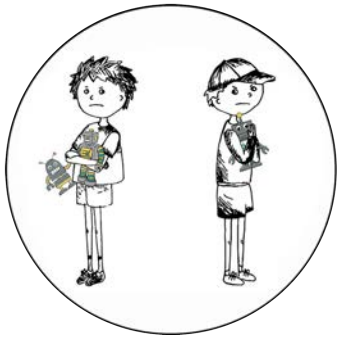


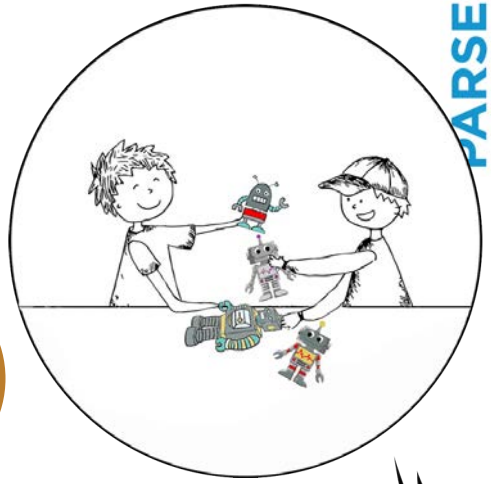
Changing culture and practice



After F. Vanclay (1992)

Changing culture and practice





PARSEC

Open Science

perception

context, situation

I shall try it!

environment

awareness

adoption

test

and trial

sustainable practice

A colleague does it

knowledge, information

resources, finance, knowledge

Are these approaches that might work?

- Provide exemplars?
- Align early with your intended repository (TRUSTed of course)?

What is a TRUSTed repository?

Principle	Guidance for Repositories
Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
User Focus	To ensure that the data management norms and expectations of target user communities are met.
Sustainability	To sustain services and preserve data holdings for the long-term.
Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

How respositories can help

"Repositories have a vital role in applying and enforcing target user community norms and standards as compliance facilitates data interoperability and reusability. Data standards...include metadata schema, data file formats, controlled vocabularies and other semantics where they exist in the user community."

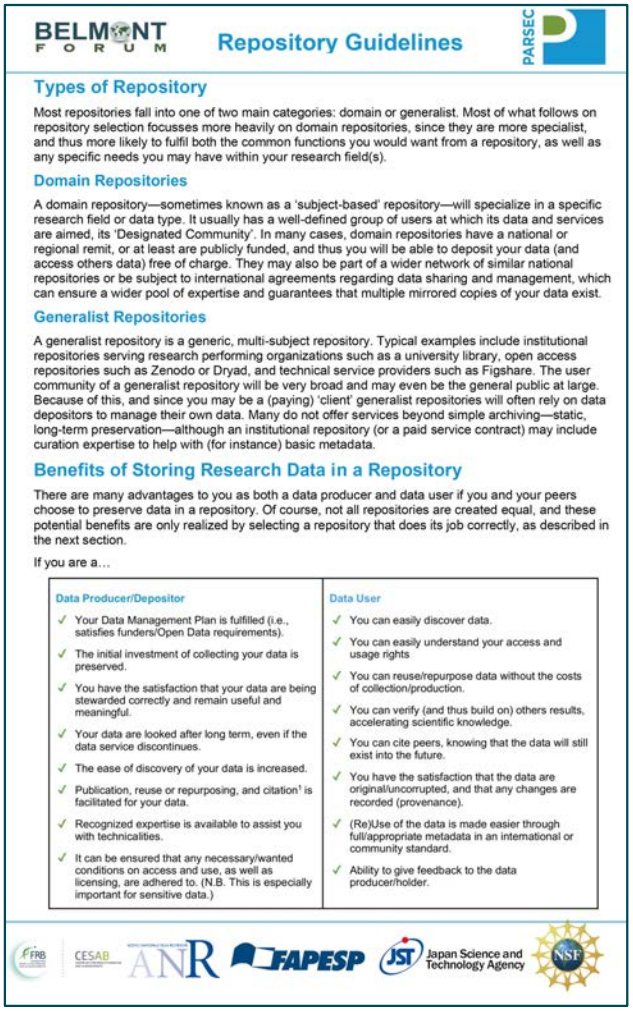
"Repositories should encourage users to fully describe data at the time of deposition and facilitate feedback on any issues with the data (e.g. quality or fitness for use) that may become apparent after the data have been made available."

Adapted from Lin et al., (2020) Scientific Data doi: 10.1038/s41597-020-0486-7

Clarification is given in our PARSEC repository guideline. (Edmunds et al. (2022). Zenodo doi: 10.5281/zenodo.6542494)

Are these approaches that might work?

- Provide exemplars?
- Align early with your intended repository (TRUSTed of course)?
- Provide educational packages?



BELMONT FORUM

Repository Guidelines

PARSEC

Types of Repository

Most repositories fall into one of two main categories: domain or generalist. Most of what follows on repository selection focusses more heavily on domain repositories, since they are more specialist, and thus more likely to fulfil both the common functions you would want from a repository, as well as any specific needs you may have within your research field(s).

Domain Repositories

A domain repository—sometimes known as a 'subject-based' repository—will specialize in a specific research field or data type. It usually has a well-defined group of users at which its data and services are aimed, its 'Designated Community'. In many cases, domain repositories have a national or regional remit, or at least are publicly funded, and thus you will be able to deposit your data (and access others data) free of charge. They may also be part of a wider network of similar national repositories or be subject to international agreements regarding data sharing and management, which can ensure a wider pool of expertise and guarantees that multiple mirrored copies of your data exist.

Generalist Repositories

A generalist repository is a generic, multi-subject repository. Typical examples include institutional repositories serving research performing organizations such as a university library, open access repositories such as Zenodo or Dryad, and technical service providers such as Figshare. The user community of a generalist repository will be very broad and may even be the general public at large. Because of this, and since you may be a (paying) 'client' generalist repositories will often rely on data depositors to manage their own data. Many do not offer services beyond simple archiving—static, long-term preservation—although an institutional repository (or a paid service contract) may include curation expertise to help with (for instance) basic metadata.

Benefits of Storing Research Data in a Repository

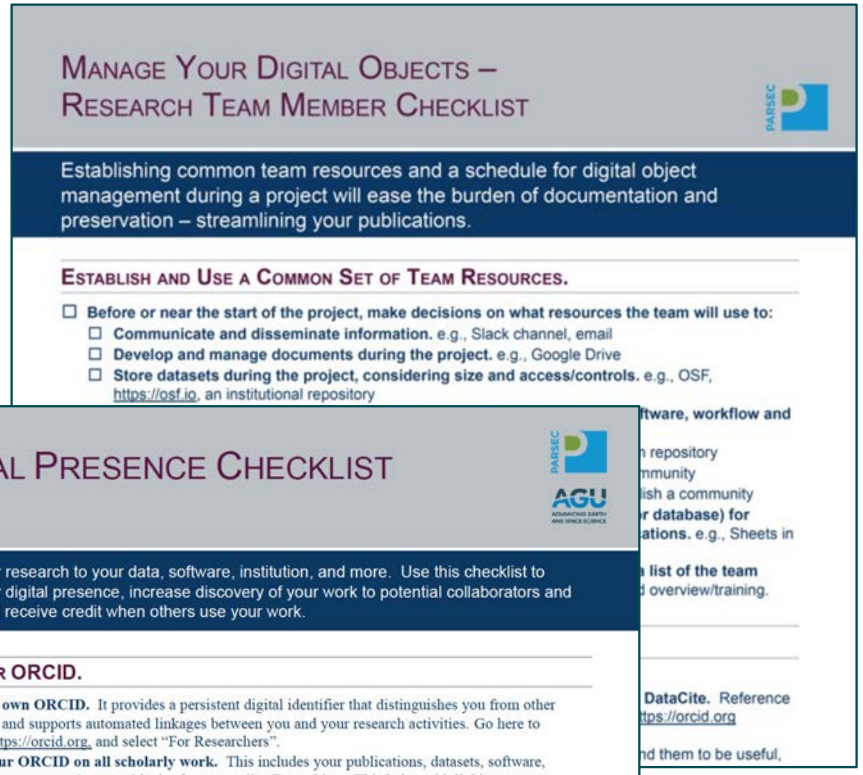
There are many advantages to you as both a data producer and data user if you and your peers choose to preserve data in a repository. Of course, not all repositories are created equal, and these potential benefits are only realized by selecting a repository that does its job correctly, as described in the next section.

If you are a…

Data Producer/Depositor	Data User
✓ Your Data Management Plan is fulfilled (i.e., satisfies funders/Open Data requirements).	✓ You can easily discover data.
✓ The initial investment of collecting your data is preserved.	✓ You can easily understand your access and usage rights
✓ You have the satisfaction that your data are being stewarded correctly and remain useful and meaningful.	✓ You can reuse/repurpose data without the costs of collection/production.
✓ Your data are looked after long term, even if the data service discontinues.	✓ You can verify (and thus build on) others results, accelerating scientific knowledge.
✓ The ease of discovery of your data is increased.	✓ You can cite peers, knowing that the data will still exist into the future.
✓ Publication, reuse or repurposing, and citation[1] is facilitated for your data.	✓ You have the satisfaction that the data are original/uncorrupted, and that any changes are recorded (provenance).
✓ Recognized expertise is available to assist you with technicalities.	✓ (Re)Use of the data is made easier through full/appropriate metadata in an international or community standard.
✓ It can be ensured that any necessary/wanted conditions on access and use, as well as licensing, are adhered to. (N.B. This is especially important for sensitive data.)	✓ Ability to give feedback to the data producer/holder.

FRB CESAB ANR FAPESP JST Japan Science and Technology Agency NSF

PARSEC has been actively creating guidelines, toolkits and a series of seminars and workshops to help users across all aspects of the research data lifecycle, including vocabularies



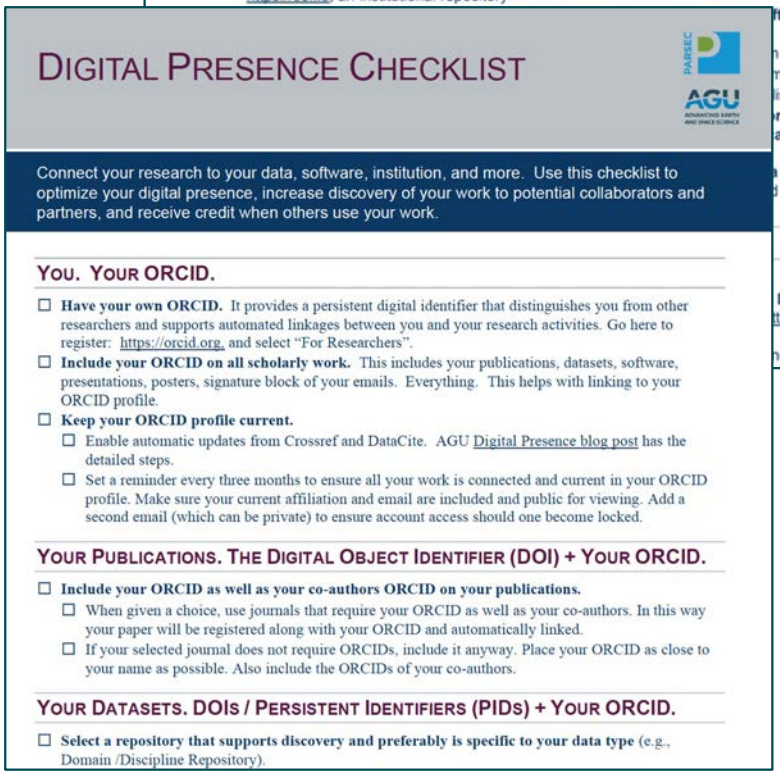
Manage Your Digital Objects – Research Team Member Checklist

PARSEC

Establishing common team resources and a schedule for digital object management during a project will ease the burden of documentation and preservation – streamlining your publications.

Establish and Use a Common Set of Team Resources.

- ☐ **Before or near the start of the project, make decisions on what resources the team will use to:**
  - ☐ **Communicate and disseminate information.** e.g., Slack channel, email
  - ☐ **Develop and manage documents during the project.** e.g., Google Drive
  - ☐ **Store datasets during the project, considering size and access/controls.** e.g., OSF, https://osf.io, an institutional repository



Digital Presence Checklist

PARSEC AGU

Connect your research to your data, software, institution, and more. Use this checklist to optimize your digital presence, increase discovery of your work to potential collaborators and partners, and receive credit when others use your work.

You. Your ORCID.

- ☐ **Have your own ORCID.** It provides a persistent digital identifier that distinguishes you from other researchers and supports automated linkages between you and your research activities. Go here to register: https://orcid.org, and select "For Researchers".
- ☐ **Include your ORCID on all scholarly work.** This includes your publications, datasets, software, presentations, posters, signature block of your emails. Everything. This helps with linking to your ORCID profile.
- ☐ **Keep your ORCID profile current.**
  - ☐ Enable automatic updates from Crossref and DataCite. AGU Digital Presence blog post has the detailed steps.
  - ☐ Set a reminder every three months to ensure all your work is connected and current in your ORCID profile. Make sure your current affiliation and email are included and public for viewing. Add a second email (which can be private) to ensure account access should one become locked.

Your Publications. The Digital Object Identifier (DOI) + Your ORCID.

- ☐ **Include your ORCID as well as your co-authors ORCID on your publications.**
  - ☐ When given a choice, use journals that require your ORCID as well as your co-authors. In this way your paper will be registered along with your ORCID and automatically linked.
  - ☐ If your selected journal does not require ORCIDs, include it anyway. Place your ORCID as close to your name as possible. Also include the ORCIDs of your co-authors.

Your Datasets. DOIs / Persistent Identifiers (PIDs) + Your ORCID.

- ☐ **Select a repository that supports discovery and preferably is specific to your data type** (e.g., Domain /Discipline Repository).

https://zenodo.org/communities/parsec/?page=1&size=20

Are these approaches that might work?

- Provide exemplars?
- Align early with your intended repository (TRUSTed of course)?
- Provide educational packages?
- Are there confounding factors?

Confounding factors to be worked around!

Many researchers, many domains, many countries, many languages, many approaches.

e.g. We, partnering with others, are developing some more options to improve the multi-linguality of data sharing (see our poster next RDA)

Confounding factors to be worked around!

Many researchers, many domains, many countries, many languages, many approaches.

e.g. We, partnering with others, are developing some more options to improve the multi-linguality of data sharing (see our poster next RDA)

Where is an accepted vocabulary that suits my work?

Despite efforts to date, there is confusion for the researcher, and in consequence the topic of a PARSEC-sponsored SciDataCon session this June will be: “Where are the vocabularies that will make environmental datasets FAIR?”

How the use of controlled vocabularies could be improved to achieve open-science outcomes.

- Provide exemplars
- Align early with your intended repository (TRUSTed of course)
- Provide educational packages
- Ensure there are work-arounds for any confounding factor.

How the use of controlled vocabularies could be improved to achieve open-science outcomes.

- Provide exemplars
- Align early with your intended repository (TRUSTed of course)
- Provide educational packages
- Ensure there are work-arounds for any confounding factor.

All of these!


NICT
国立研究開発法人
情報通信研究機構
National Institute of Information and Communications Technology
RIHN
Inter-University Research Institute Corporation
National Institutes for the Humanities
Research Institute for Humanity and Nature
TOKYO METROPOLITAN UNIVERSITY
UNIVERSITÉ TOULOUSE III PAUL SABATIER
UNIVERSITE MONTPELLIER
USP
SciELO
IRD
Institut de Recherche pour le Développement
FRANCE
cnrs
DataCite
FIND, ACCESS, AND REUSE DATA
Thankyou!
marbec
marine biodiversity exploitation & conservation
RDA
RESEARCH DATA ALLIANCE
ADVANCING EARTH AND SPACE SCIENCE
SIP
BELMONT FORUM
JST
AGENCE NATIONALE DE LA RECHERCHE
ANR
NSF
WORLD DATA SYSTEM
PARSEC
FRB
FONDATION POUR LA RECHERCHE SUR LA BIODIVERSITÉ
CESAB
CENTRE FOR THE SYNTHESIS AND ANALYSIS OF BIODIVERSITY
Powell Center
FAPESP
FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
erinha
European Research Infrastructure on Highly Pathogenic Agents
NCI
AUSTRALIA
PROVIDING AUSTRALIAN RESEARCHERS WITH WORLD-CLASS HIGH-END COMPUTING SERVICES