Standardised licence classification: an approach born from use-cases and lived experience

RDA 14th Plenary, Helsinki

24th October 2019

Graham Parton, Sam Pepler, Kate Winfield

This work is licensed under a Creative Commons Attribution 4.0 International License.

Quick caveat!

I’m not a legal or licence expert!

But what follows is from lived experience running an evolving, 25+ year old archive.

What is CEDA?

Centre for Environmental Data Analysis

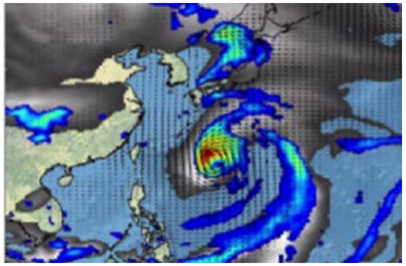
Mission: To provide data and information services for environmental science

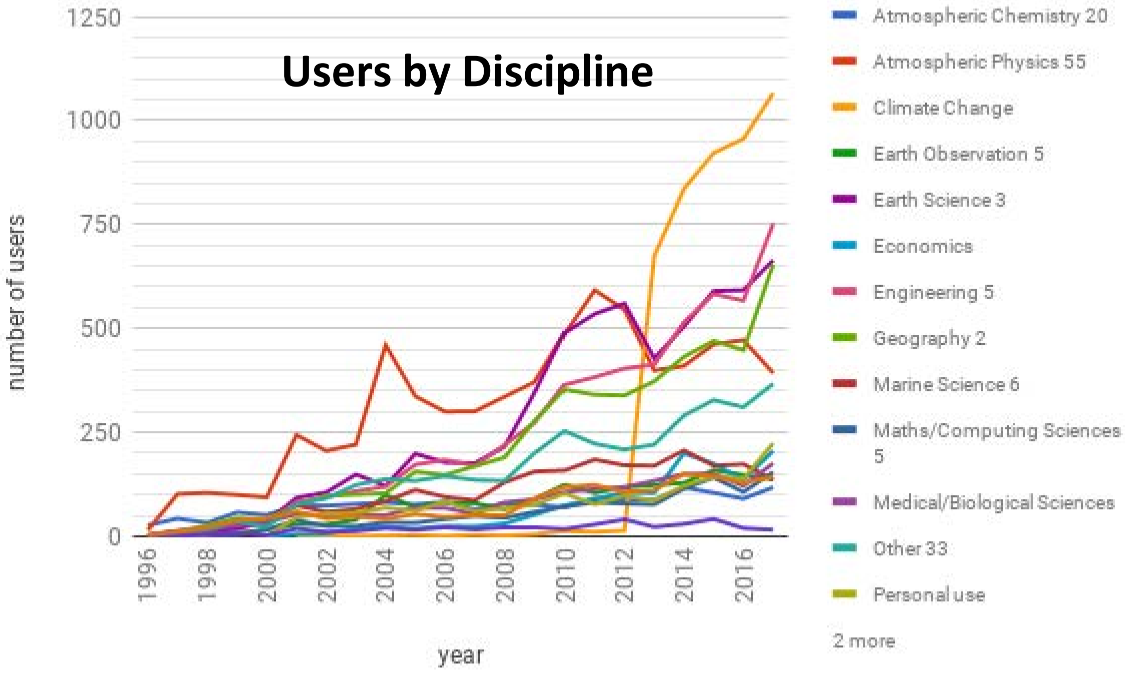
~30 staff, mixture of data scientists and software engineers in STFC RAL Space

- Expertise in:
  - *Earth observation*
  - *Climate modelling*
  - *Aircraft measurements*
  - *Data standards*
  - *Data services*
  - *… and much more!*

CEDA Data

Data Type	Data Volume (Petabytes)
Earth Observation	11.1
Atmospheric Science	5.2
Total	16.3 PB





- 5622 datasets
- In 576 dataset collections
- Covering ~ 224 million files
- > 57,600 registered users (+ unregistered users)

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Data Licences

- Have used over 100 data licences over 25 year period
- Still have 80+ licences in operation
- Drop mainly due to move towards standard licences (CC, UK Open Gov licences, Closed-Use generic licences)
- Many bespoke, project-orientated licences of varying quality remain

Licence related questions

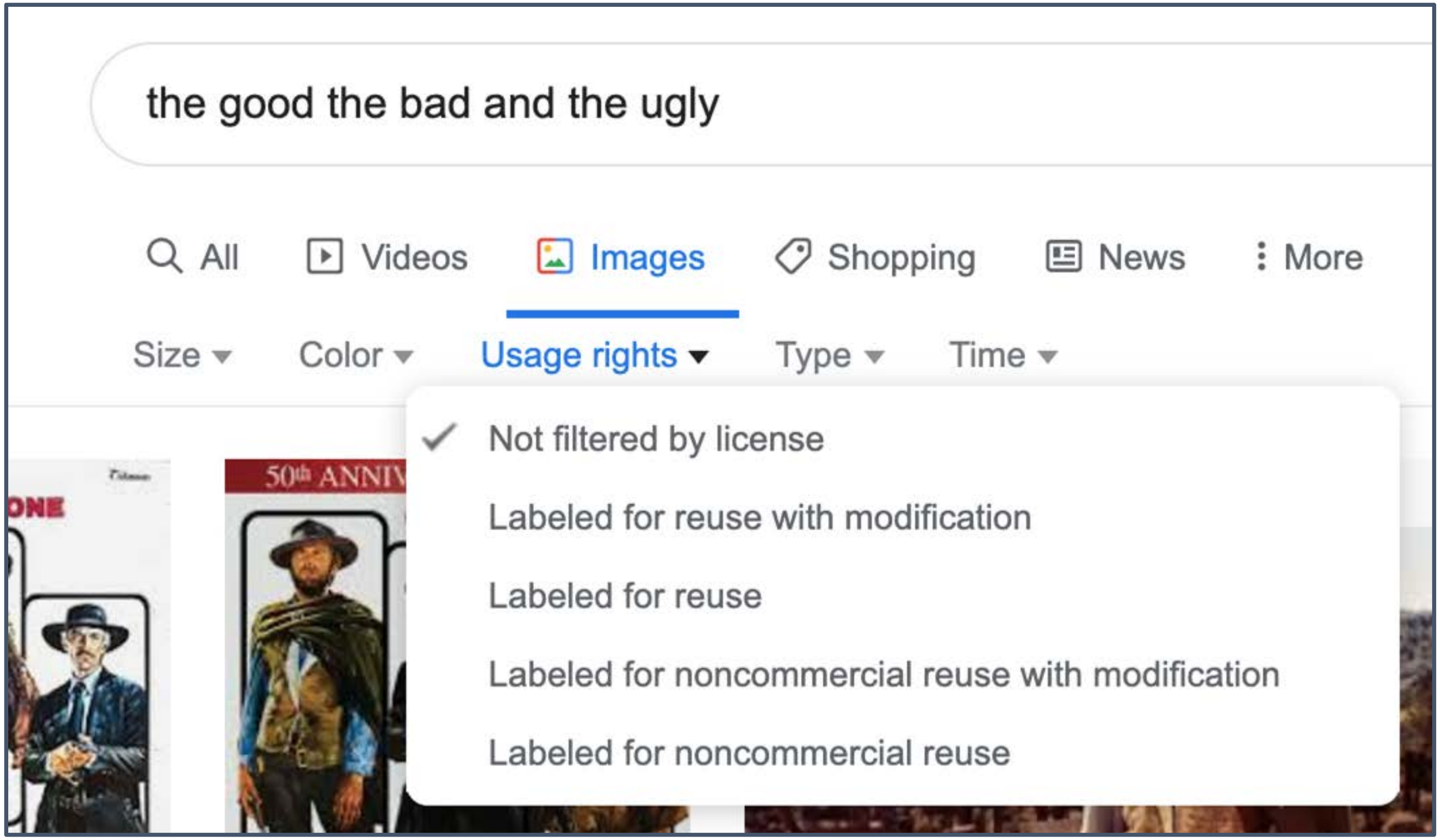
- Users want to know which datasets are available for:
  - Commercial use
  - Policy use (both government and NGOs)
  - Personal use
  - Teaching use
- Stakeholders want to know what ‘impact’ we have
  - E.g. number of commercial datasets

Both use-cases centre around filtering datasets by permitted *use*.

Note - this is *use-types* **not** *user-types*, as academic users may do commercial or policy work for example.

Ideal

Some sort of usage-rights filter akin to those on Google Image… but with greater nuance



the good the bad and the ugly

✓ Not filtered by license

Labeled for reuse with modification

Labeled for reuse

Labeled for noncommercial reuse with modification

Labeled for noncommercial reuse

2014 - new catalogue and 1st licence review

CEDA released new ISO standard data catalogue, but :

- no ability to search by permitted use
- No list of commercially available datasets

Set out to see if we could classify licences to support these requirements

What did we find?

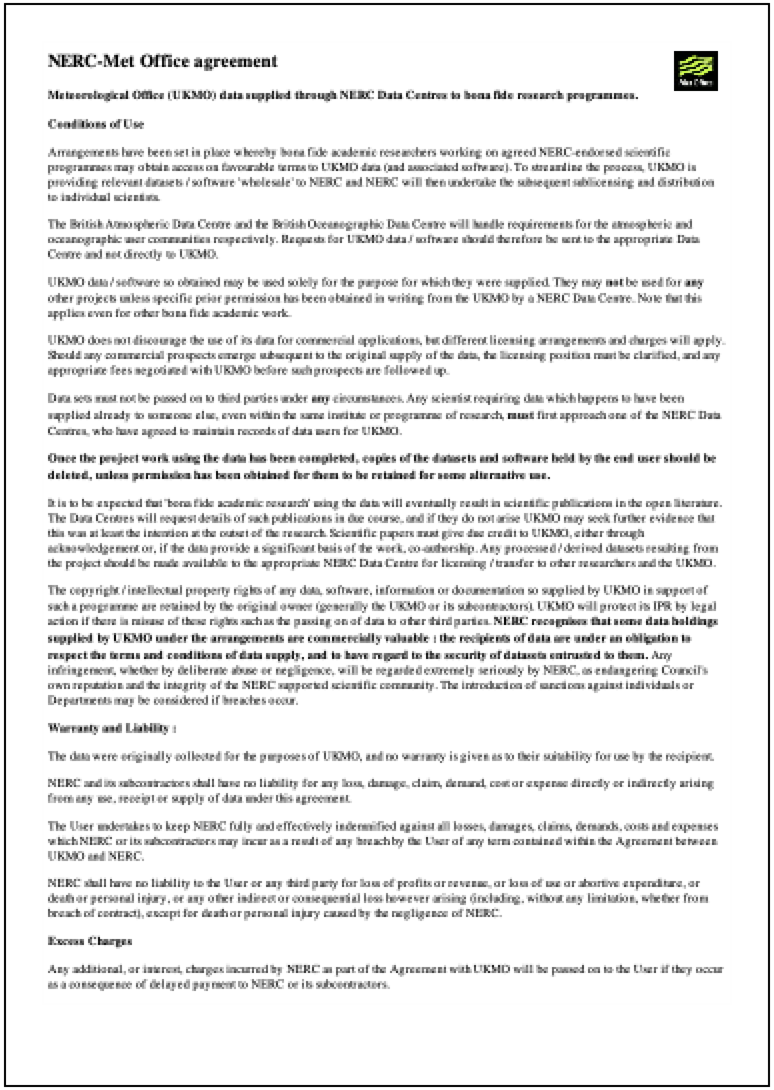
- Quality of licence varies enormously
- Range of licences from very specific to generic licences
- Identified broad categories of permitted-use

Essentially, a bit of a licencing wild-west has existed over the years.

What we found confirmed assumptions from > 12 years experience of dataset application administration

CEDA licences: the **Good**

- Generic or organisation level licences
- Well structured
- Clearly defined use-scope



NERC-Met Office agreement

Meteorological Office (UKMO) data supplied through NERC Data Centres to bona fide research programmes.

Conditions of Use

Arrangements have been set in place whereby bona fide academic researchers working on agreed NERC-endorsed scientific programmes may obtain access on favourable terms to UKMD data (and associated software). To streamline the process, UKMO is providing relevant datasets / software 'wholesale' to NERC and NERC will then undertake the subsequent sublicensing and distribution to individual scientists.

The British Atmospheric Data Centre and the British Oceanographic Data Centre will handle requirements for the atmospheric and oceanographic user communities respectively. Requests for UKMO data / software should therefore be sent to the appropriate Data Centre and not directly to UKMO.

UKMO data / software so obtained may be used solely for the purpose for which they were supplied. They may **not** be used for **any** other projects unless specific prior permission has been obtained in writing from the UKMO by a NERC Data Centre. Note that this applies even for other bona fide academic work.

UKMO does not discourage the use of its data for commercial applications, but different licensing arrangements and charges will apply. Should any commercial prospects emerge subsequent to the original supply of the data, the licensing position must be clarified, and any appropriate fees negotiated with UKMO before such prospects are followed up.

Data sets must not be passed on to third parties under **any** circumstances. Any scientist requiring data which happens to have been supplied already to someone else, even within the same institute or programme of research, **must** first approach one of the NERC Data Centres, who have agreed to maintain records of data users for UKMO.

Once the project work using the data has been completed, copies of the datasets and software held by the end user should be deleted, unless permission has been obtained for them to be retained for some alternative use.

It is to be expected that 'bona fide academic research' using the data will eventually result in scientific publications in the open literature. The Data Centres will request details of such publications in due course, and if they do not arise UKMO may seek further evidence that this was at least the intention at the outset of the research. Scientific papers must give due credit to UKMO, either through acknowledgement or, if the data provide a significant basis of the work, co-authorship. Any processed / derived datasets resulting from the project should be made available to the appropriate NERC Data Centre for licensing / transfer to other researchers and the UKMO.

The copyright / intellectual property rights of any data, software, information or documentation so supplied by UKMO in support of such a programme are retained by the original owner (generally the UKMO or its subcontractors). UKMO will protect its IPR by legal action if there is misuse of these rights such as the passing on of data to other third parties. **NERC recognises that some data holdings supplied by UKMO under the arrangements are commercially valuable : the recipients of data are under an obligation to respect the terms and conditions of data supply, and to have regard to the security of datasets entrusted to them.** Any infringement, whether by deliberate abuse or negligence, will be regarded extremely seriously by NERC, as endangering Council's own reputation and the integrity of the NERC supported scientific community. The introduction of sanctions against individuals or Departments may be considered if breaches occur.

Warranty and Liability :

The data were originally collected for the purposes of UKMO, and no warranty is given as to their suitability for use by the recipient.

NERC and its subcontractors shall have no liability for any loss, damage, claim, demand, cost or expense directly or indirectly arising from any use, receipt or supply of data under this agreement.

The User undertakes to keep NERC fully and effectively indemnified against all losses, damages, claims, demands, costs and expenses which NERC or its subcontractors may incur as a result of any breach by the User of any term contained within the Agreement between UKMO and NERC.

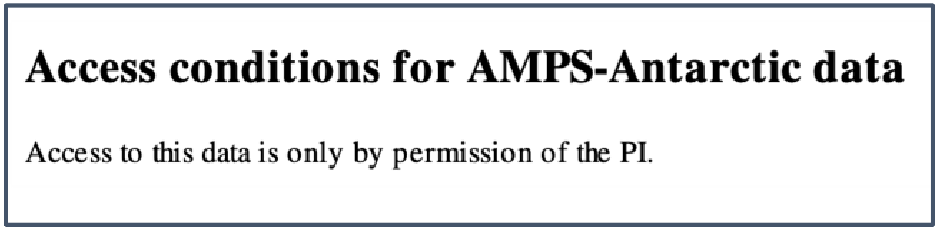
NERC shall have no liability to the User or any third party for loss of profits or revenue, or loss of use or abortive expenditure, or death or personal injury, or any other indirect or consequential loss however arising (including, without any limitation, whether from breach of contract), except for death or personal injury caused by the negligence of NERC.

Excess Charges

Any additional, or interest, charges incurred by NERC as part of the Agreement with UKMD will be passed on to the User if they occur as a consequence of delayed payment to NERC or its subcontractors.

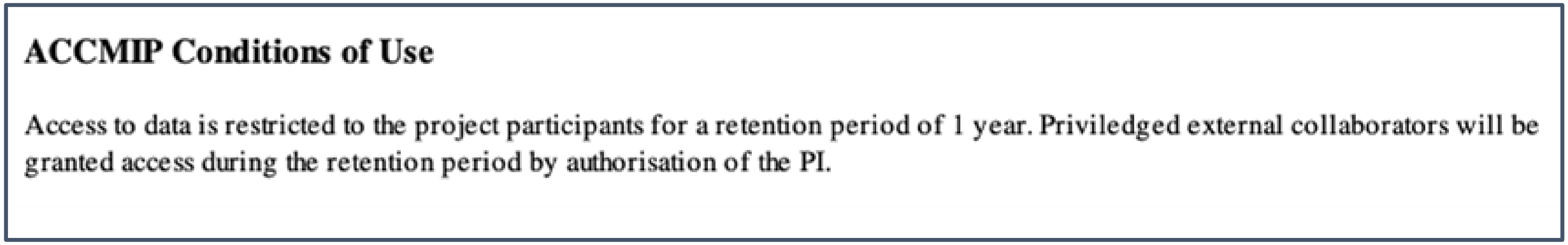
CEDA licences: the Good, the **Bad**

- Not really a licence!
- Very little content
- Don't indicate what use you can make of the data!



Access conditions for AMPS-Antarctic data

Access to this data is only by permission of the PI.

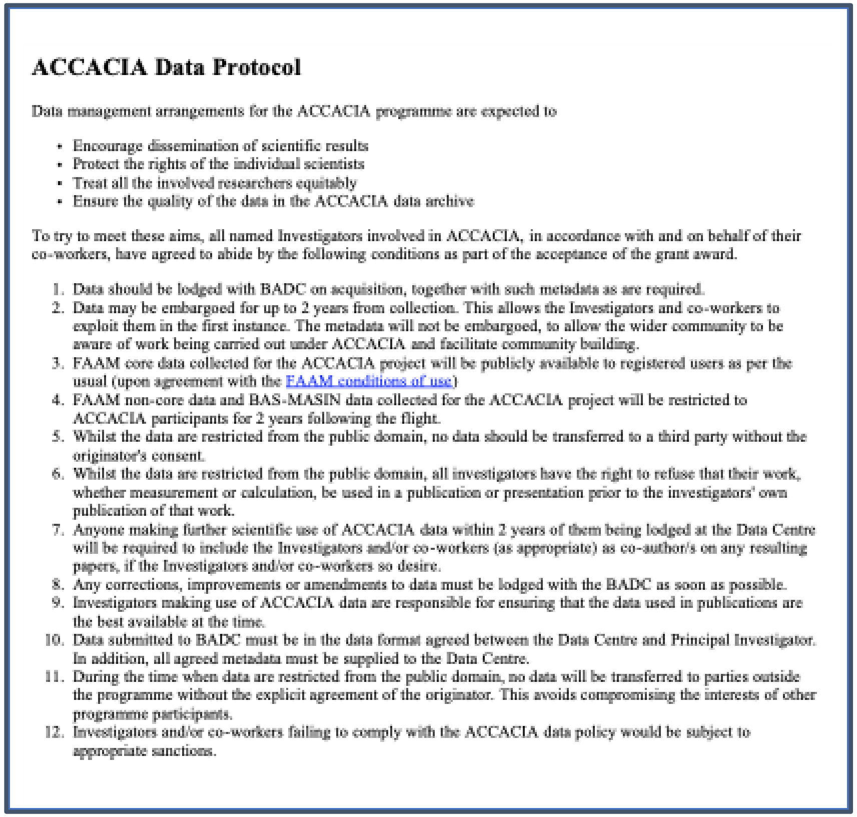


ACCMIP Conditions of Use

Access to data is restricted to the project participants for a retention period of 1 year. Priviledged external collaborators will be granted access during the retention period by authorisation of the PI.

CEDA licences: the Good, the Bad and the **Ugly**!

- An amalgam of different content:
- Data management details
- Some licence content
- Hard to determine permitted-use



ACCACIA Data Protocol

Data management arrangements for the ACCACIA programme are expected to

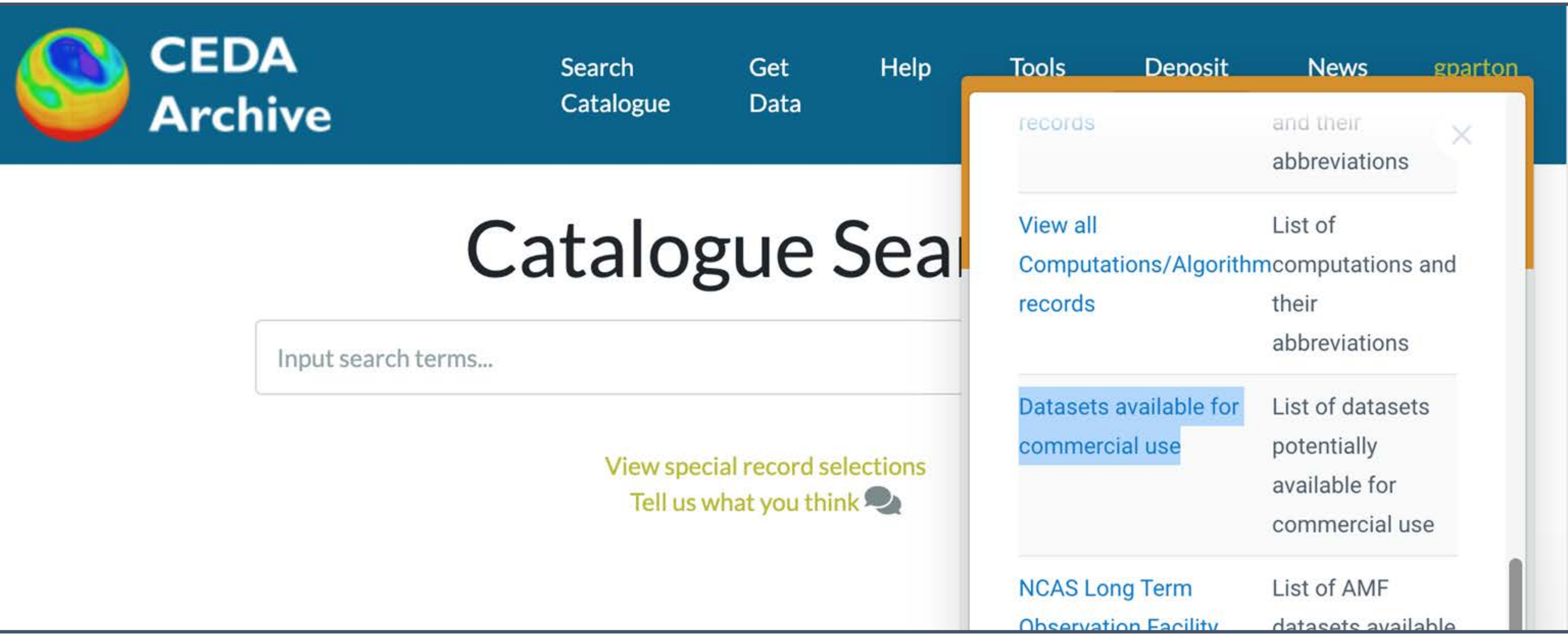
- Encourage dissemination of scientific results
- Protect the rights of the individual scientists
- Treat all the involved researchers equitably
- Ensure the quality of the data in the ACCACIA data archive

To try to meet these aims, all named Investigators involved in ACCACIA, in accordance with and on behalf of their co-workers, have agreed to abide by the following conditions as part of the acceptance of the grant award.

1. Data should be lodged with BADC on acquisition, together with such metadata as are required.
2. Data may be embargoed for up to 2 years from collection. This allows the Investigators and co-workers to exploit them in the first instance. The metadata will not be embargoed, to allow the wider community to be aware of work being carried out under ACCACIA and facilitate community building.
3. FAAM core data collected for the ACCACIA project will be publicly available to registered users as per the usual (upon agreement with the FAAM conditions of use)
4. FAAM non-core data and BAS-MASIN data collected for the ACCACIA project will be restricted to ACCACIA participants for 2 years following the flight.
5. Whilst the data are restricted from the public domain, no data should be transferred to a third party without the originator's consent.
6. Whilst the data are restricted from the public domain, all investigators have the right to refuse that their work, whether measurement or calculation, be used in a publication or presentation prior to the investigators' own publication of that work.
7. Anyone making further scientific use of ACCACIA data within 2 years of them being lodged at the Data Centre will be required to include the Investigators and/or co-workers (as appropriate) as co-author/s on any resulting papers, if the Investigators and/or co-workers so desire.
8. Any corrections, improvements or amendments to data must be lodged with the BADC as soon as possible.
9. Investigators making use of ACCACIA data are responsible for ensuring that the data used in publications are the best available at the time.
10. Data submitted to BADC must be in the data format agreed between the Data Centre and Principal Investigator. In addition, all agreed metadata must be supplied to the Data Centre.
11. During the time when data are restricted from the public domain, no data will be transferred to parties outside the programme without the explicit agreement of the originator. This avoids compromising the interests of other programme participants.
12. Investigators and/or co-workers failing to comply with the ACCACIA data policy would be subject to appropriate sanctions.

CEDA licences: making use of what we found

- Commercial datasets now listed
- But still not a search facet



Catalogue Sea

Input search terms...

View special record selections
Tell us what you think

records	and their abbreviations
View all Computations/Algorithm records	List of computations and their abbreviations
Datasets available for commercial use	List of datasets potentially available for commercial use
NCAS Long Term Observation Facility	List of AMF datasets available

2019 - 2nd licence classification review and refinement

Classification scheme revised and codified to support the following use cases:

1. Provide a search mechanism for stakeholders to find datasets for a given use type
2. A quick short hand to show users the permitted-use of a dataset in catalogue record listing and views
3. Aid internal review to find meaningful licences (and avoid re-use of old, unsuitable licences)
4. Ensure that licences have been check if they are legally sound
5. Aid licence selection for DMP purposes

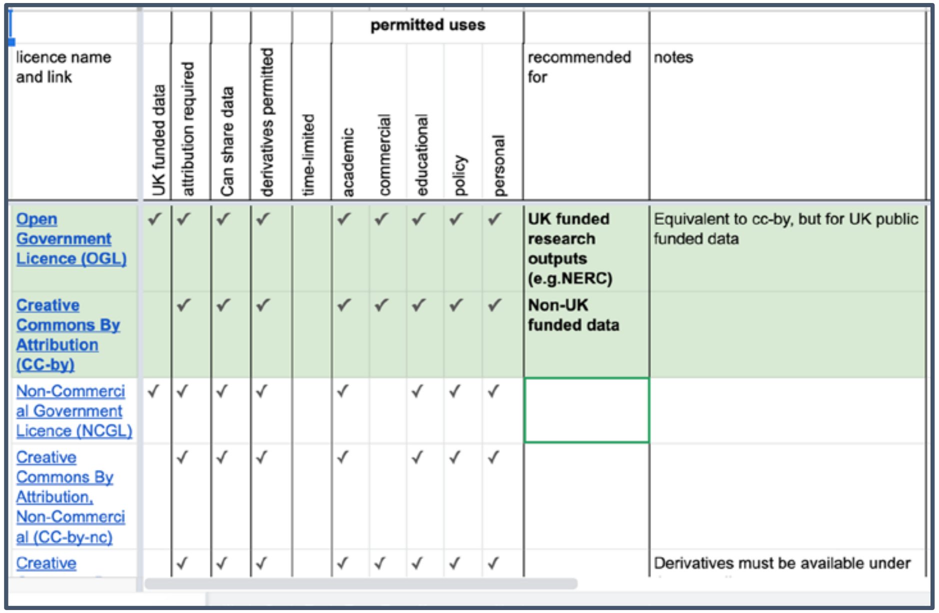
(6. Aid quick access application assessment against licences)

2019 - 2nd licence classification review and refinement

Code	Definition	Search Facets	Licence tagging	Internal audit	Legality audit	Licence selection
any	any use is permitted	√	√			√
academic	may be used for academic research, resulting in results being publically available	√	√			√
commercial	may be used for financial gain	√	√			√
educational	may be used for educational purposes	√	√			√
policy	may be used internally within organisations to aid development of policies and procedures, including governmental use	√	√			√
personal	may be used for personal, non-commercial, use	√	√			√
specific	specifically defined permitted use given, see licence for details		√	√		√
unclear	the permitted use is not clearly defined within the terms of the licence		√	√		
unstated	no permitted use has been stated by the licence		√	√		
unclassified	no use review yet undertaken		√	√		
legal	has been reviewed by a legal expert and found to be acceptable			√	√	√
notlegal	has been reviewed by a legal expert and found to be legally unsound				√	√

2019 - Making use of the classification scheme

The review process has helped CEDA follow much better licence practice!



licence name and link	UK funded data	attribution required	Can share data	derivatives permitted	time-limited	permitted uses academic	commercial	educational	policy	personal	recommended for	notes
Open Government Licence (OGL)	✓	✓	✓	✓		✓	✓	✓	✓	✓	**UK funded research outputs (e.g.NERC)**	Equivalent to cc-by, but for UK public funded data
Creative Commons By Attribution (CC-by)		✓	✓	✓		✓	✓	✓	✓	✓	**Non-UK funded data**	
Non-Commercial Government Licence (NCGL)	✓	✓	✓	✓		✓		✓	✓	✓		
Creative Commons By Attribution, Non-Commercial (CC-by-nc)		✓	✓	✓		✓		✓	✓	✓		
Creative		✓	✓	✓		✓	✓	✓	✓	✓		Derivatives must be available under

Clear structure for:

- choosing one of 12 main generic licences (including a set of new generic licences covering ‘Closed-use’ and ‘Restricted-Use’ requirements)
- to choose from and guidelines to assess bespoke licences *where these are essential*

CEDA Implementation plan

- Classification scheme applied to licences (done)
- Storing licence permitted-use classification scheme in catalogue (next few months)
- Setting up search facet in CEDA catalogue (by end of year?)
- Review how to annotate export version of records (e.g. put this in ISO19115 ‘usage limitation’ text, but this lacks any encoded way to support external interpretation)

Summary: we have a workable *local* solution only.

Benefits of standardisation

- External portals able to provide search licence facets
- Agreed definitions
- Universally understood by users
- Could look at encoding other licence aspects (e.g. need for attribution, geographic limits, time-limits)

One approach:

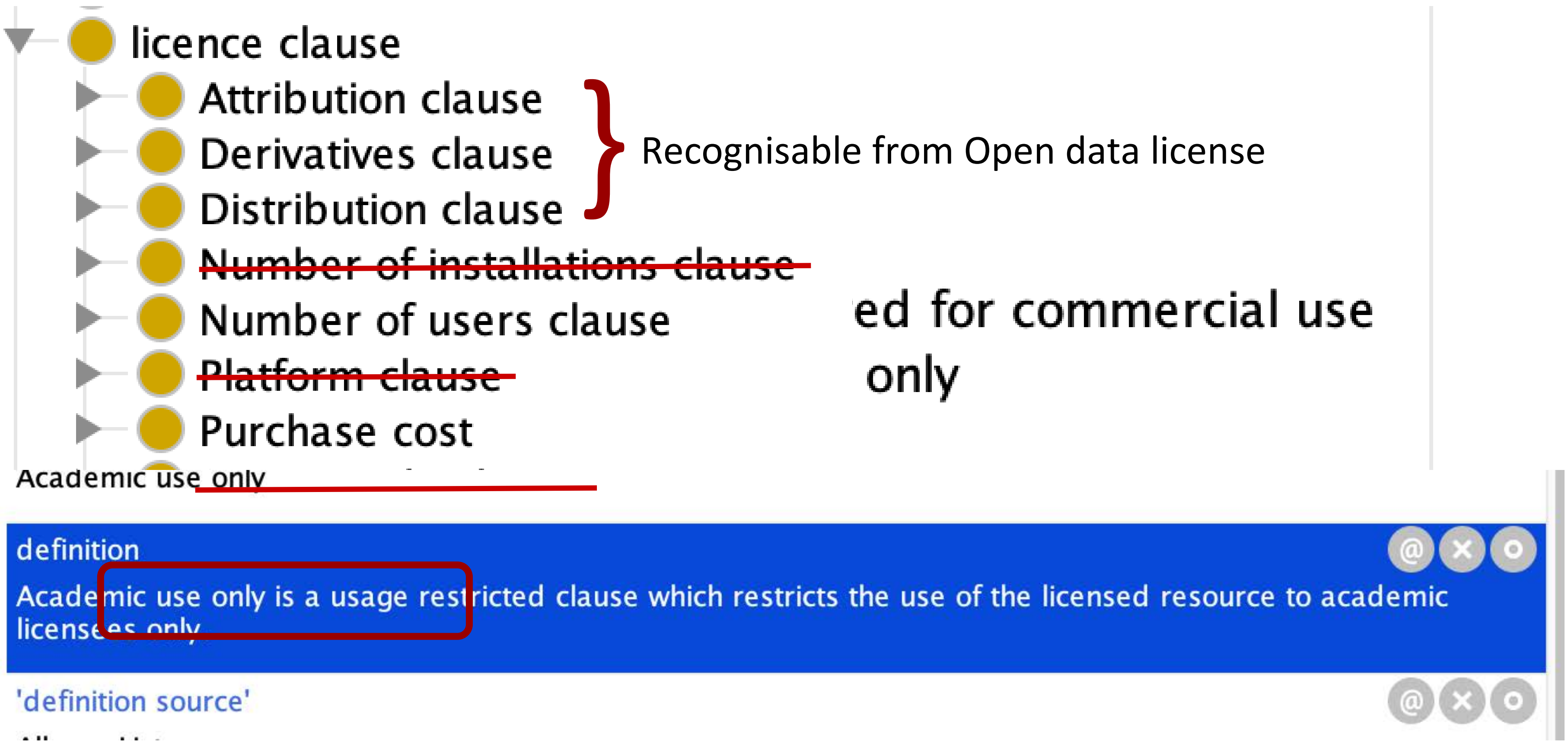
Build on licence clause work of Software Ontology:

For more details see:

https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-5-25

Though for recent work see:
https://github.com/allysonlister/swo/blob/master/LicenceHierarchy.md - version 1.7 was released on Monday!

Software Ontology Licence Clauses



- licence clause
  - Attribution clause
  - Derivatives clause
  - Distribution clause
  - ~~Number of installations clause~~
  - Number of users clause
  - ~~Platform clause~~
  - Purchase cost

Recognisable from Open data license

ed for commercial use only

Academic use only

definition

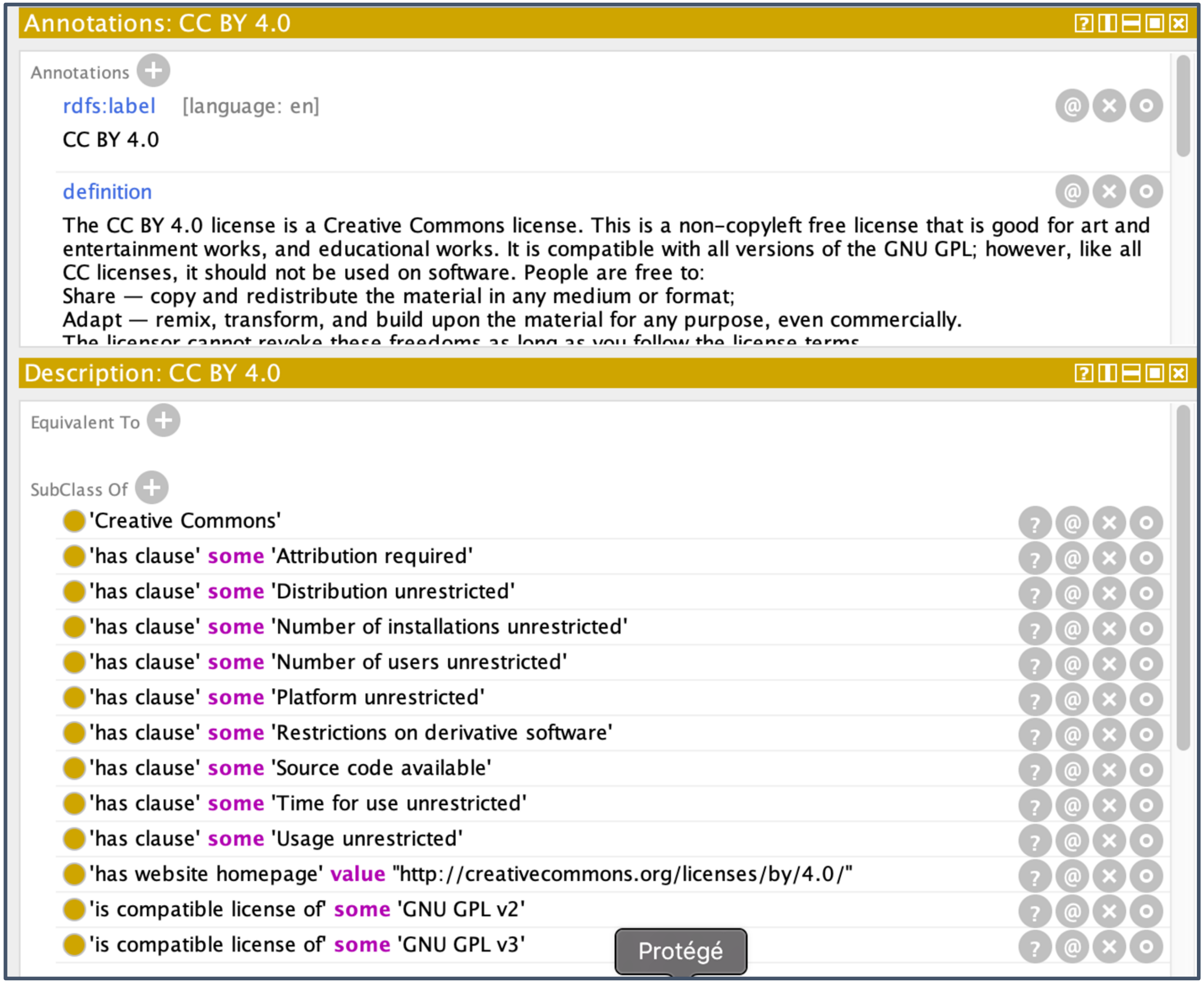
Academic use only is a usage restricted clause which restricts the use of the licensed resource to academic licensees only

'definition source'

I.e. this is user-type focused, not use-focused

This only has some uses covered, more needed (.. but that’s OK as it’s based on Open World Assumptions)

Software Ontology Licence Clauses



Annotations: CC BY 4.0

Annotations

rdfs:label [language: en]

CC BY 4.0

definition

The CC BY 4.0 license is a Creative Commons license. This is a non-copyleft free license that is good for art and entertainment works, and educational works. It is compatible with all versions of the GNU GPL; however, like all CC licenses, it should not be used on software. People are free to:
Share — copy and redistribute the material in any medium or format;
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms

Description: CC BY 4.0

Equivalent To

SubClass Of

- 'Creative Commons'
- 'has clause' **some** 'Attribution required'
- 'has clause' **some** 'Distribution unrestricted'
- 'has clause' **some** 'Number of installations unrestricted'
- 'has clause' **some** 'Number of users unrestricted'
- 'has clause' **some** 'Platform unrestricted'
- 'has clause' **some** 'Restrictions on derivative software'
- 'has clause' **some** 'Source code available'
- 'has clause' **some** 'Time for use unrestricted'
- 'has clause' **some** 'Usage unrestricted'
- 'has website homepage' **value** "http://creativecommons.org/licenses/by/4.0/"
- 'is compatible license of' **some** 'GNU GPL v2'
- 'is compatible license of' **some** 'GNU GPL v3'

Feedback from poster session at RDA 14

Summary of comments from poster session mentioned:

- Split the code list into a multidimensional model/structure
- This codifying of licences would aid Virtual Research Environments (data and tools) to bring together different resources by their permitted use
- Would aid selection of resources for hackathons
- Discussing how to address licences identified as having issues, suggestion was to look at examples where copyright issues have been resolved
- Carr's work on data use agreements elements presented at a previous RDA could be relevant
- Other licence aspects to consider:
    - What about CARE as well as FAIR?

Question - does the code list cover artistic/performance use?

Poster available at:

https://docs.google.com/presentation/d/1p6l0YAPcnf16k0uPmspd6xohuDYGjQbufpV8ieMRVIo/edit?usp=sharing

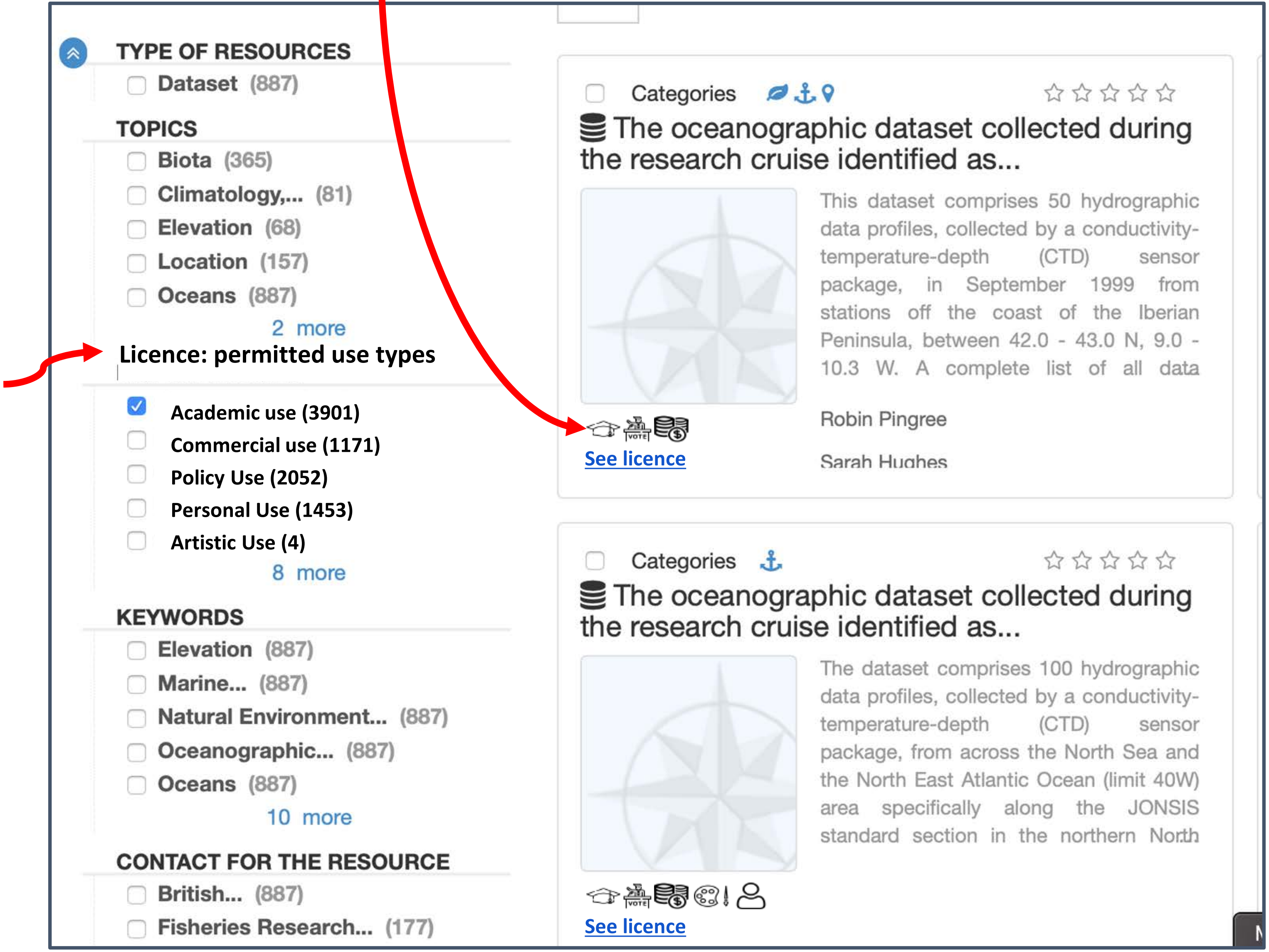
Next steps?

THE END

The dream?

Meaningful permitted-usage icons for quick reference

Licence clause filters available



TYPE OF RESOURCES

□ Dataset (887)

TOPICS

□ Biota (365)

□ Climatology,... (81)

□ Elevation (68)

□ Location (157)

□ Oceans (887)

2 more

Licence: permitted use types

☑ Academic use (3901)

□ Commercial use (1171)

□ Policy Use (2052)

□ Personal Use (1453)

□ Artistic Use (4)

8 more

KEYWORDS

□ Elevation (887)

□ Marine... (887)

□ Natural Environment... (887)

□ Oceanographic... (887)

□ Oceans (887)

10 more

CONTACT FOR THE RESOURCE

□ British... (887)

□ Fisheries Research... (177)

□ Categories ☆☆☆☆☆

The oceanographic dataset collected during the research cruise identified as...

This dataset comprises 50 hydrographic data profiles, collected by a conductivity-temperature-depth (CTD) sensor package, in September 1999 from stations off the coast of the Iberian Peninsula, between 42.0 - 43.0 N, 9.0 - 10.3 W. A complete list of all data

Robin Pingree

Sarah Hughes

See licence

□ Categories ☆☆☆☆☆

The oceanographic dataset collected during the research cruise identified as...

The dataset comprises 100 hydrographic data profiles, collected by a conductivity-temperature-depth (CTD) sensor package, from across the North Sea and the North East Atlantic Ocean (limit 40W) area specifically along the JONSIS standard section in the northern North

See licence

Icons made by ***turkkub*** *from www.flaticon.com*