Workshop - Choose the Right Rights, Use the Data Right

Draft Summary

These are informal notes taken on the day and we apologise for any inaccuracies or omissions.

Overview

On 10th February, the University of Glasgow and Jisc held an all-day workshop on licencing research datasets.

The purpose of the workshop was to update the community on how their input to previous workshops has been used and discuss current issues and potential solutions in licencing datasets.

The project aimed to provide guidance for:

- Creators of data - understand and choose most appropriate licence to release their dataset under
- Consumers of datasets - understand what they can and cannot do with data
- The Jisc Research Data Shared Service project which is now a live service known as Jisc Research Hub.

Four Guides were produced:

Introduction to Ownership of Rights in Research Data http://eprints.gla.ac.uk/171314/

Making Research Data Available

http://eprints.gla.ac.uk/171315/

Choosing a Licence for Research Data

http://eprints.gla.ac.uk/171316/

FAQ: Using Research Data

http://eprints.gla.ac.uk/171317/

Guides and tools like these need to be updated to be of use in the future.

Key Recommendations from this Workshop

It is clear that there is a need for ongoing advice on dataset licencing rights and terminology

We recommend that the community continue to discuss and share best practices around dataset licencing. This might include:

- Annual review of the guides and other tools by volunteers from the community
- Finding a way to fund legal expertise to update the guides

- A community created glossary of licencing terms perhaps via the CASRAI framework https://casrai.org/
- Contacting funders to see how they can support this work associated with the outputs of their funding
- Contact other initiatives such as the Research Data Alliance, and the Digital Preservation Coalition to see if there is synergy. Joint action may result in better support for the community
- Explore if further lay person explanation of licencing can be progressed. This might be similar to or based on the Centre for Environmental Data Analysis (CEDA) example below.

Demo of Jisc Research Hub

http://eprints.gla.ac.uk/211156/

Jisc initially asked Universities what they needed to respond to funder requirements.

They piloted a repository which is now live and available.

This was set up with minimum required metadata requirements to reduce barriers to data deposit.

https://www.jisc.ac.uk/open-research-hub

Datasets don't always have clean single licences which causes problems. Licences can be mixed.   This complicates the user experience - multiple tick boxes can be off-putting or lead to metadata being poorly recorded.   The Research Hub development focussed heavily on user experience.

The goal was to obtain rich metadata and not be 'strict and shouty about it'

There are five minimum compulsory bits of metadata and any other fields are made as easy as possible to complete.

Allows 'don't know' so that after initial record creation someone else who does know can help the user e.g. to choose the right licence if they are unsure what the licence choices mean.

Can set generic or individual licence for each file.

Can save a draft and add additional information later.

Terms and conditions are defined by the research organisation e.g. to make sure users are not uploading copyrighted material.

The workflow can be set up to automatically run files through a digital preservation process.

Questions:

If we need to keep data for 10 years does it need to be online or can it be offline?

Data is often kept on spinning disk. Sometimes it is slower to retrieve from online file storage services such as Amazon Glacier, or from tape. As long as there is a method to retrieve in a reasonable timeframe then funders should be happy.

Sensitive data can be restricted in the repository. The metadata can be open without the data being available. Some people might say that at least metadata should be open to follow the ethos of openness requested by funders and good research practice.

There is a list of specific licences - can people have other licences?

Research Organisations can have any licences they want.

What if you wanted access to a closed file?

This depends on how depositor has classified access e.g. recommend a named role at an organisation rather than a person as people may move on and not be contactable.

Is the info already published elsewhere? Does it then discourage deposit to avoid 2 canonical options? (Answer was not know on the day)

Consider the risk of loss.

Consider if a location is trustworthy and has longevity.

If researchers upload, is there an institutional approval?

Yes.

Legal Update

We developed our guides based on questions and clarifications requested by attendees at our previous workshops.

However there is a necessary level of complexity that cannot be solved with reading. Reading the guides will not make you a lawyer.

In the current climate there are new laws being discussed and Brexit both of which may influence dataset licencing.

There has been a huge package of EU copyright reform recently.

- Copyright in the Single Digital Market - much press coverage and discussion

- Filtering Obligations for Online Content Sharing Providers (probably little impact on us as mainly directed at video e.g. YouTube)

Text and data mining (TDM) for non-commercial and research have exceptions in UK and EU law.

Deadline for implementation of the TDM EU directive is beyond the transition period for Brexit so the UK government is not bound by it. They may still choose to adopt this law and enshrine in UK law or as part of an agreement with the EU. It was recently declared they do not plan to translate to UK law.

The UK will probably still have an exception limited to non-commercial uses. But could remove that as recommended in the Hargreaves review. https://www.gov.uk/government/publications/digital-opportunity-review-of-intellectual-property-and-growth

A further EU directive with strong and opposing opinions is the Public Sector Directive. This relates to data held by public bodies such as City Councils. The most recent update was 2019. Again it may be beyond Brexit transition when this is implemented so there is no obligation for the UK to adopt it. The current law is implemented in the UK allowing re-use of public body information once it has been made accessible. Re-useability by default. Regulated by the UK government.

Member states are expected to have national policies for open access to data compatible with FAIR data principles

https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf

This will apply to research organisations and to funders. Non derivative options would not be compatible. **Mandatory open access to research data.**

Consider how collaborations with EU will work in future if a combination of UK and EU law.

What if we collaborate with industry? Might need a case made as to why not to share.

Deposit Agreements and Workflows

Some repositories mediate data deposits for the creator. They may not resourced to check in detail the data that is being ingested therefore want to have a method to ensure it is clear who has which obligations. Different research organisations do different things - forms, paper sign off, sometimes agreement is not explicit. Sometimes it is part of the repository and sometimes separate e.g. on a web page. E.g. the depositor must declare they have not infringed copyright.

The attendees were given a copy of the draft agreement for Glasgow and asked to discuss at their tables.
The feedback will be used to update the Glasgow agreement.

We discussed whether creative commons or some other agnostic organisation could take on hosting template agreements.

Workshop - Lay Person flags, Graham Parton, Centre for Environmental Data Analysis (CEDA)

A copy of the presentation is available with a link to a recorded version on the first slide:

https://docs.google.com/presentation/d/1QmyYJolMJB51Tp94GR8TaEMZOZNfj4RbYdEN4tpYlp0/edit#slide=id.g50d8b18f131632ba_319

Have had over 100 types of licence over 25 years.

Why are they not using generic open licences?

- Third party content
- Onward sharing is sometimes not desirable
- Inherited licence in some case
- Some data providers don't want to be open

What are data available for - commercial, policy, personal, teaching use?

Focus is on 'use type' not 'use**r** type' - users may be of one 'type' (e.g. academic) but their *use* of data may be something else (e.g. use data for a commercial project).

Looking at something like google image rights easy to understand options for dataset licences.

There are:

Good - well structure, generic
Bad - not really a licence, little content, dont explain how you can use the data
Ugly - hard to determine permitted use

Licences.

CEDA wanted to help users find data that might be usable by them for different purposes and aid licence selection.

2019 licence classification scheme included
- use type e.g.. commercial, personal

- Level of clarity e.g. specific,unclear
- Legality - legal or not legal

Structure provided to new data providers to help them choose the right licence.

Is there a standard way out there already?

The dream - select type of resource, select uses that it can be used for - a licence will then be suggested.

The Software Ontology https://www.ebi.ac.uk/ols/ontologies/swo

Some aspects not necessary e.g. platform clause but could be ignored. This is looking mainly at restrictions and mainly academic. Have already codified standard licences such as CC-BY.

Open Digital Rights Language https://www.w3.org/TR/odrl/ is a way to make web 2.0 understandable 'policy' documents.

Licence, parties, clauses, for a specified asset.

Some tools are being built around this such as convertors to machine readable versions.

CEDA next steps:

Internal implementation of and community cooperation.

Take home messages:
- Licencing hard, but...
- Licencing is important
- Better options now for generic licencing
- Need machine readable to aid discovery and use

Is the work published anywhere? Not so far.

Need for use cases - there are some EU parties looking at verifying this.

Workshop - Requirements

The attendees worked in groups to identify current issues in dataset licencing. 3 key areas were identified and discussed further.

Data Sharing Agreements

These seem to be increasing in volume. How do we advise users?

One comment - you cannot 'own' data therefore don't need a data sharing agreement. The fact that it is 6c outside is data - no-one owns that bit of data.

If you have to have a data sharing agreement consider:

Scope out the data - is it personal? non-personal? Possibly to protect by intellectual property rights?

Use existing tools and laws

Don't conflate GDPR with other data
More definitions and guidance would be useful.

Licence Stacking

I want to combine multiple datasets from different sources in my research. These datasets have (very) different licences.

- How do I know whether I can do the work I have planned? Are the licences compatible?

- How do I licence my output from this project?

Seek advice from a body such as Jisc who could create a tool or advice to help decide what licences can be combined and what licence to apply to the new dataset.

Default to open licences if possible.

Try the OpenMinted licence tool that shows what you can do with combinations of licences.
https://openminted.github.io/releases/license-matrix/

Future proofing Licence Protection

What aspects of licencing do we need to future proof?

- Research students own their data. Employees usually do not.
- What facilities are in place to keep track of datasets?
- How long do licences last? Can we change the licence if the dataset is expanded to include information that should be published under another licence?
- Are licences comprehensive enough to manage orphan data?

Versioning was felt to be important here. New versions could be posted with new licences.

Other Issues Raised

How should we licence physical samples e.g. rocks? Suggestion to contact organisations that might do this already e.g. National Geoscience Data Centre.

If I write software how can I be sure it is used appropriately?

Is there any move towards licence convergence?

What is the smallest amount of information a researcher needs to choose a licence correctly?

Complexity of data sharing with different partners

How can this licencing work be sustained? The law changes and advice/guidance needs to be kept up to date.

Where can I find guidance on legitimate exceptions to data sharing? e.g. REF Conditions

How can we progress the work piloted by CEDA in other data catalogues? (e.g. integrate classification scheme and promote interoperability between data catalogues)

Copyright in images e.g. catalogue images of paintings

Orphan data - reverts to organisational ownership?

How to enforce a licence if data is used on the other side of the world?

If I obtained some of my data via data mining what licence do I apply when I share it?

What is the 'right' way to add a licence to a dataset? e.g. in metadata only, in read-me file, both, other?

What advice would you give to researchers to help make the process easier for everyone?

Those in this room are engaged….but many in the community are not. How best can the importance of/need for licences be propagated/disseminated?

Next Steps

Sustainability -

Knowledge base ask questions and answer questions?

Regular community review of guidance e.g. virtually, community.

Who pays? Who takes responsibility?

Academic environment geared round 4* outputs for Research Excellence Framework so cannot sustain by expecting academic lawyers to provide time.

Who else can help? Create a bridge between technology transfer offices and libraries but need a position or intern to work on specific issues. Someone still needs to pay and how is that fair to a specific organisation that does?

Is the scope just HEI - no wider research community - can funders help?

Does RDA have a group?

Action: Follow up via mail list.

Acknowledgements

Donaldson, Graham Parton and Thomas Margoni who brought their own expertise and perspective to the event.