Choose the Right Rights, Use the Data Right

Draft Summary

Overview

On Thursday 2nd November, the [University of Glasgow Research Information Management Services Team](), [Jisc](), and [CREATe]() held an all-day workshop on licencing research datasets.

The purpose of the workshop was to draw out issues around licencing of research datasets with a focus on identifying whether clarifications in terminology and guidance would be useful.

Key Recommendations

Jisc, CREATe and the University of Glasgow will discuss these recommendations some of which will be actioned as part of the current piece of work that initiated this workshop. Updates on activity will be provided to those who registered for the workshop and via https://datasetlicencing.wordpress.com/

- Create and maintain high level lay guidance and process diagrams including benefits of using specific licence types.

- Provide good examples of appropriate and inappropriate use of licences that researchers and users can access.

- Consider discipline specific guidance and examples e.g. music and creative media have specific complexities.

- Provide guidance on licencing physical media and samples.

- Make training on how to use licences appropriately available to all stakeholders. Include how to assign a licence and how to check the licence attached to any data that is used.

- Recommend attribution at organisation level and provide practical guidance on how to do this.

- Recommend sign up to [Concordat on Open Research Data]()

- Encourage funders to provide guidance on preferred licences.

- Provide guidance on working with commercial partners and producing or using datasets.

- Provide a glossary of terms and plain language translation possibly in collaboration with [CASRAI]().

- Provide tools for licence selection and automatically applying licence metadata to multiple files and embedding licence information in file headers.

- Machine actionable licences e.g. MS Word can include CC-BY in metadata of document.

- DMP Online update to provide guide to channel researchers to consider all sources of data they might use and what the licences allow.

Presentations and Activities

Case Study 1
Zosia Beckles talked us through the University of Bristol's approach to database licencing. To address some of the omissions from early Creative Commons licences, the University of Bristol adopted the Non-commercial Government Licence 2.0 as its default licence for databases, although it is in process of changing this to CC-BY 4.0. Zosia talked attendees through the University of Bristol deposit process and the mechanisms by which researchers are introduced to database licencing. She highlighted some challenges for for example, where multiple licences might be possible or preferred and the balance that needs to be struck between appropriate licencing and a user-friendly streamlined system that encourages researchers to make their datasets open access. The general feeling at the Workshop was that the University of Bristol were providing more information about database licencing than some other institutions, although Zosia stressed that there were difficulties with engagement and that, with the exception of a highly-engaged software engineering field, most researchers did not engage with the process unless required to for publication.

Licence Overview

Thomas Margoni from CREATe http://www.create.ac.uk/ presented some of the legal challenges arising in relation to database licencing. He summarised the legal protection available for research datasets; this is a complex area. Creative content within a database (for example, literary pieces of writing, or visual data) might attract copyright protection, but less 'creative' content will not automatically be protected. The structure of a database itself might qualify for copyright protection, but the content will only be protected if it meets the criteria for the 'sui generis database right' (SGDR) which requires 'substantial investment' in obtaining, verifying or presenting the contents. A particular discussion point was the construction of some research databases - many are simply a collection of files in folders, however a database is legally defined as a collection of individually accessible data 'arranged in a systematic or methodical way'. Thomas then discussed the most common type of licences often used for databases - CC-BY 4.0, CC-BY 3.0, CC-BY-SA, CC-BY-ND-NC, CC0, OGL 3.0, and ODC BY and the advantages and disadvantages of applying these to research databases. While CC-BY 4.0 is often the default in the sector, Thomas highlighted that the SGDR does not actually include a right to be attributed, so by applying this licence non original databases, users are applying a condition that that the law does not require. Thomas then summarised some of the work that the OpenMinTeD project has done in relation to licence compatibility in the field of text data mining

Licence Interpretation Exercise

Attendees then split up into groups and looked at four common licences used for datasets: CC-BY 4.0, CC0, derivatives of CC-BY and OGL 3.0. The issues with these licences ranged from the 'hard sell' of the non-attribution of CC0 to academics, difficulties in understanding the terminologies and wording used across different licences, to correctly attaching licencing metadata to uploaded files and avoiding intermediary liability for infringement of third party rights. Making datasets Open Access presents resource implications. Administrators might not be able to look at every dataset individually, or face nonchalance from researchers about the type of licence they need their dataset to have. A fear of reverting to a default position of not making databases available due to uncertainty was also common. Attendees were often keen to stress that knowledge and need varied across academic disciplines, with those in software often keen to write their own licences and those in Science, Technology, Engineering and Mathematics (STEM) subjects concerned about commercial use. The [Public Sector Information (PSI) Directive](#) in relation to the applicability of the [OGL 3.0](#) was also discussed.

Case Study 2

After lunch, Andrew McHugh from the [Urban Big Data Centre](#) (UBDC), an ESRC funded data centre based at the University of Glasgow, presented on the contribution UBDC has made to helping researchers access datasets across a broad range of fields in addition to generating its own data. UBDC has strong relationships with major data owners and a growing base of users across the UK. Some of the challenges faced by UBDC are common across the research management field - incompatibilities with business models, concerns about reputation where data is perceived to be of low quality, concerns from data owners that their data may be used in research that casts them in a poor light, management of multiple licences and lack of resources. Andrew discussed some of the issues with personal data governed by data protection laws, and how anonymisation of data can present difficulties in making datasets open, using the [integrated Multimedia City Data](#) project as an example. UBDC has assisted many researchers negotiate access to controlled data with e.g. NHS Scotland.

Discussion

The workshop closed with further group discussion and feedback on database licencing options and the key issues and challenges in this area. For some groups, use of third party data presented the most challenges and support was needed to standardise resources and terminology, and understand how data can be sourced and used from the beginning of creating a Data Management Plan. Looking beyond the present, would it be possible for machines to collate, interpret and action licences? For others, the interoperability of licences and the possibility of combination licences is an area of difficulty, particularly when dealing with funders' requirements in the commercial sphere. What was clear was that different academic fields encounter different challenges in making their data open access and sector specific guidance is needed. The question of physical samples within data was raised; these are often subject to Material Transfer Agreements to transfer them to other users on request, but do research managers have other responsibilities to make them available under Open Access? The difficulties of selling non-attribution licences such as CC0 to academics created a discussion about the need for best practice guidelines on attribution. Finally staff training and best practice in uploading, attaching metadata and avoiding intermediary liability for the institution was discussed and attendees agreed much more guidance was needed.

Feedback from the workshop was positive with all who commented saying it was a very worthwhile meeting. It was valuable to have the CREATe expertise. Comments included:

'Very stimulating - woke up lots of dusty bits of the brain!'

'Generated a lot of questions and action points'

'Just never had the opportunity to explore it at such length before'

Another comment reminded us that we have come a long way already in terms of data management and any improvement is good - we should not worry too much about getting everything perfect right now.

Acknowledgements