Legal *Inter*-operability & *Intra*-operability of Research Data:

*The Case of the Research Compendium*

Gail Clement,
RDA/CODATA Legal Interop IG Co-Chair
Head of Research Services, Caltech Library

https://orcid.org/0000-0001-5494-4806 | @repositorian
So, what is a Research Compendium?
1.3 Research compendia

The term *research compendium* was coined by Gentleman and Lang (2007) who “introduce[d] the concept of a compendium as both a container for the different elements that make up the document and its computations (i.e. text, code, data,...), and as a means for distributing, managing and updating the collection.” According to Marwick, Boettiger, and Mullen (2018) it provides “a standard and easily recognisable way for organising the digital materials of a research project to enable other researchers to inspect, reproduce, and extend the research”. This standard may differ between scien-
Also known as ....

Jupyter notebooks

Interactive Notebooks

R Projects
R Notebooks

Whole Tale Vision

- Living publication (data + code + environment)
- Facilitate reproducibility
- Encourage investigation of results making it easy to recreate the environment the result was created in
Also known as

Living Paper

eLife's new computationally reproducible article will allow users to modify figures
Characteristics of the Research Compendium
- Data rich digital research object
- Combines text and markup, underlying data, code for analysis, figures/plots
- Multiple files representing diverse formats and types of digital objects
- Electronically packaged into a compound object for digital distribution

Western blot of c-Myc reactivation in tetracycline-repressible system and total RNA expression during c-Myc re-activation.

Target of replication: A priori experimental protocols, materials, and confirmatory analysis plans were peer reviewed, pre-registered, and published by cff.re.

This component includes all primary and QC data for the replication of Lin et al. 2012. This experiment included the Western blot of c-Myc reactivation in tetracycline-repressible system and total RNA expression during c-Myc re-activation.
Why worry about the Legal Interoperability of a Research Compendium?
Why worry about licensing of Research Compendium

- Gaining wider interest among research authors: Early adopter publishers are attracting buzz with their experiments.
- Observable confusion and misunderstanding.
- Hard to find good examples of licensing done right.
- Not-so-hard to find examples of worrisome licensing practices.
Licensing for research compendia?

As I was creating the repo and was confronted with assigning a license, I was stumped on how best to license this project. In @benmarwick, @cboettig, and @lincoln paper the discuss licensing and follow the suggestions in https://web.stanford.edu/~vcs/papers/ERROLSI03092009.pdf. Essentially the different parts of a research compendia should have different licenses. For example, the manuscript and figures could use CC-BY, the code MIT, and the data CC0.

My question is more on the nuts and bolts best practice of capturing the fact that a compendia would be released under the multiple licenses. Should I:

- List License: CC-BY, MIT, CC0 in the DESCRIPTION
- Explain in a README
- Both
- Something else?

I've dug around and can't find an implementation of this multiple license concept for compendia. Thoughts most welcome!
Question 1. Won’t one omnibus license cover the whole aggregation of objects?
Example 1:

Jupyter Notebook (and related files) zipped as one record in CaltechData Repository and used locally with relevant installed dependencies
<table>
<thead>
<tr>
<th>License statement #1</th>
</tr>
</thead>
<tbody>
<tr>
<td>cc-by</td>
</tr>
</tbody>
</table>

Cite this record as:
Belliveau, N., Barnes, S., Ireland, W., & Phillips, R. (2018). Processed data, code, and Jupiter notebooks from "Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria" (Version 1.0.6) [Data set]. CaltechDATA. https://doi.org/10.22002/d1.960 or choose a different citation style.
The downloaded and unzipped files look like this on the desktop

License statement #2:

MIT License
Copyright (c) 2017 Rob Phillips group @ California Institute of Technology
Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.
Supplemental Information Section B  Analysis of library diversity

(c) 2017 the authors. This work is licensed under a Creative Commons Attribution License CC-BY 4.0. All code contained herein is licensed under an MIT license.

In [1]: import scipy ndimage
   # Our numerical workhorses
   import numpy as np
   import pandas as pd
   import scipy as sp

   # Import the project utils
   import sys
   sys.path.insert(0, './code/)
   import NB_sortseq_utils as utils

   # Import matplotlib stuff for plotting
   import matplotlib.pyplot as plt
   import matplotlib.cm as cm
   from IPython.core.pylabtools import figsize

   # Seaborn, useful for graphics
In a computational narrative, which part is the “text” (cc-by license) and which part is the “code”? (MIT license)? A machine may not see the difference!
Example 2:

Reproducible article published in Elife with a Binder-ized manifestation of the article to enable interactivity online by anyone

Getting started with our reproducible article

Our reproducible article demo showcases just some of the functionality that RDS tools will permit, and it's intended as an easy starting point for exploring the technology. Here's what you can do:

• Look out for the round blue 'R script'-labelled buttons below Figure 1. Click it to reveal the code that generated that figure.
• Edit that code, and press shift-enter to re-run it.
• Observe the results in the figure in real time.

Future iterations will enable fully downloadable datasets and table data, more interactive figure types, and the ability to download a pre-packaged DAR source file to make it much easier to fully replicate the whole reproducible manuscript in your local environment.

You can see more of the RDS' potential in action by taking a look at this demo of Stencilia, one of the platforms behind the project.

https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article
Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc

L Michelle Lewis, Meredith C Edwards, Zachary R Meyers, C Conover Talbot Jr, Haiping Hao, David Blum, Reproducibility Project: Cancer Biology
University of Georgia, Bioexpression and Fermentation Facility, United States, Johns Hopkins University, Deep Sequencing and Microarray Core Facility, United States

REPLICATION STUDY Jan 9, 2016

Abstract

As part of the Reproducibility Project: Cancer Biology, we published a Registered Report (Blum et al., 2015), that described how we intended to replicate selected experiments from the paper Transcriptional amplification in tumor cells with elevated c-Myc (Lin et al., 2012). Here we report the results. We found overexpression of c-Myc increased total levels of RNA in P493-6 Burkitt's lymphoma cells; however, while the effect was in the same direction as the original study (Figure 3E; Lin et al., 2012), statistical significance and the size of the effect varied between the original study and the two different lots of serum tested in this replication. Digital gene expression analysis for a set of genes was also performed on P493-6 cells before and after c-Myc overexpression. Transcripts from genes that were active before c-Myc induction increased in expression following c-Myc overexpression (similar to the original study (Figure 3F; Lin et al., 2012)). Transcripts from genes that were silent before c-Myc induction also increased in expression following c-Myc overexpression, while the original study concluded elevated c-Myc had no effect on silent genes (Figure 3F; Lin et al., 2012). Treating the data as paired, we found a statistically significant increase in gene expression...
This is a Reproducible document. See the original article or source.

Providing runtime environment. This may take up to a few minutes.
Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc

I. Michelle Lewis, Meredith C. Edwards, Zachary R. Meyers, C. Conover Talbot Jr., Haiping Hao, David Blum, Reproducibility Project: Cancer Biology

As part of the Reproducibility Project: Cancer Biology, we published a Registered Report (Blum et al., 2013), that described how we intended to replicate selected experiments from the paper Transcriptional amplification in tumor cells with elevated c-Myc (Lin et al., 2012). Here we report the results. We found overexpression of c-Myc increased total levels of RNA in Pagg-3 Burkitt's lymphoma cells; however, the effect was in the same direction as the original study (Figure 3E; Lin et al., 2012), statistical significance and the size of the effect varied between the original study and the two different lots of serum tested in this replication. Digital gene expression analysis for a set of genes was also performed on Pagg-3 cells before and after c-Myc overexpression. Transcripts from genes that were active before c-Myc induction increased in expression following c-Myc overexpression, similar to the original study (Figure 3F; Lin et al., 2012). Transcripts from genes that were silent before c-Myc induction also increased in expression following c-Myc overexpression, while the original study concluded elevated c-Myc had no effect on silent genes (Figure 3F; Lin et al., 2012). Treating the data as paired, we found a statistically significant increase in gene expression for both active and silent genes upon c-Myc induction, with the change in gene expression greater for active genes compared to silent genes. Finally, we report meta-
NOTE: Below is a reproducible version of Figure 1B. You can inspect the code, make changes and run the code by pressing SHIFT+ENTER. The data used can be downloaded here.

```r
library(Rmisc)
library(ggplot2)
library(cowplot)

# names raw data from protocol 2 from csv file
data2 <- read.csv("article/Study_48_Protocol_2_Data.csv", header=T, sep="", )

# creates new column calculating RNA in 100ul
data2$RNA.100ul <- data2$Average.RNA.Concentration*100

## calculates RNA per cell
data2$RNA.per.cell <- data2$RNA.100ul/data2$Total.Cells.Harvested

# calculates RNA per 1000 cells
data2$value <- data2$RNA.per.cell*1000
```
Digital gene expression analysis.

PAS-a cells grown in the presence of tetracycline (tet) for 72 hr for repression of the conditional pyrcite construct were switched into Tet-free growth medium to induce c-Myc expression. Cells ...

Figure 2—source data 1
List of Reporter codesets and gene expression values.

https://doi.org/10.7554/eLife.30274.003

Download elife-30274-fg2-data1-v2.csv
Questions about Licensing Practices for the RC

1. Does a license in the metadata record inherit all the way through the file directory to apply to all files?
2. Can we devise a top level rights statement in the metadata record that links to a more detailed list of licenses associated with each file?
3. What is the legal status of a compendium with conflicting licenses?
Questions about Licensing Practices for the RC

- Does the license need to appear in each manifestation of the paper? (HTML, PDF, and the interactive version)
- What does ‘code’ mean in the context of computational narrative where the text itself may be computed?
- Other ????
Please join our IG to discuss many other legal questions around FAIR licensing of Research Compendia!

https://rd-alliance.org/groups/rdacodata-legal-interoperability-ig.html