Sep 15, 2019

University Published Paper Reproducibility Assessment Protocol - For Teachers

Scott Edmunds[1], Jesse Xiao[2], Hongling Zhou[3]

[1]GigaScience/BGI Hong Kong/Bauhinia Genome, [2]GigaScience/BGI Hong Kong, [3]GigaScience/BGI

1 Works for me dx.doi.org/10.17504/protocols.io.6x8hfrw

BGI

Hongling Zhou

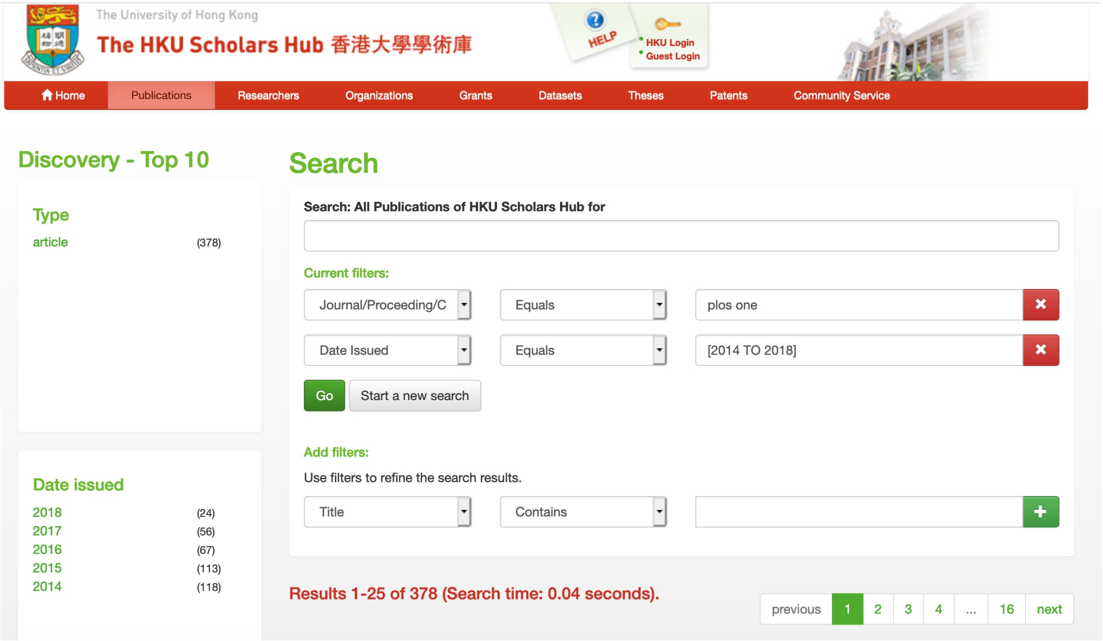
ABSTRACT

This protocol is for teachers to set up a data curation exercise for their students. Using a universities IR or CRIS to select papers that their researchers have published, and then assign them to students to see if the data availability meets the national, funder, university and journal policies. The students carrying out this assessment have a separate related protocol they need to follow.

1 **Pick journals of interest**
Use the Universities institutional repository (IR) or CRIS (current research information system) to look for journal(s) of interest to determine the number of papers that need to be assigned per student.

1.1 Carrying out this experiment as part of the HKU MLIM Data Curation MLIM course we used HKU scholars hub (http://hub.hku.hk/), chosing papers in the Open Access journals *PLOS One* and *Scientific Reports* (search results link) for the students to check, filtering based on the following:

Journal/Proceeding/Conference: [*PLOS One*] or [*Scientific Reports*]
Date Issued: [2016 TO 2018]



Discovery - Top 10

Type

article	(378)

Date issued

2018	(24)
2017	(56)
2016	(67)
2015	(113)
2014	(118)

Search

Search: All Publications of HKU Scholars Hub for

Current filters:

Journal/Proceeding/C Equals plos one

Date Issued Equals [2014 TO 2018]

Add filters:

Use filters to refine the search results.

Title Contains

Results 1-25 of 378 (Search time: 0.04 seconds).

Screenshot of typical institutional repository search demonstrating the fields/filters used.

1.2 Alternatively, if the repository is built upon DSpace this information can be webscraped out using this open source python script:

https://github.com/jessesiu/hku_scholars_hub

The requirements are: python 3+ and installation of: Install BeautifulSoup bs4, urlopen, xlrd,xlwt,csv. You will need to input the journal name and timeline you'd like to search, and the result file will save as test.txt in the current directory. The output includes the repository identifier (handle, etc.), the manuscript DOI, and it also parses out the contents of any "Data Availability" section in the XML.

HKU Scholarhub URL	Manuscript DOI URL	Data Availability Comment
http://hub.hku.hk/handle/10722/241589	http://dx.doi.org/10.1371%2Fjournal.pone.0169095	All relevant data are within the paper and its Supporting Information files.

Example of the output from the python script webscraping DSpace institutional repositories.

2 Assignment of papers to students

Divide out and assign a suitable number of papers per student for them to assess for data availability in the data curation exercise.

2.1 In our case we assigned 4-5 papers per student and pasted the details into a googledoc spreadsheet with the handle/DOI/URL of each paper they needed to assess. We recommended the students should spend about 5-10 minutes checking each paper. Instructions for each student to carry out the exercise are given in the related student protocol, and students were told to input their results into the appropriate sections of the spreadsheet for the teacher to check. See the example spreadsheet in the table below that can be adapted:

Assigned to student	**Identifier of paper (URL/DOI/Handle)**	**Is there data presented in the paper?**	**Is there external data, and if so what is the link/accession?**	**Is all the data in the paper available?**	**Comments**	**For teacher: If "data available on request", do they respond when contacted (try twice & include dates)?**
[student name]	e.g. http://hub.hku.hk/handle/xxxxx/xxxxxxxx	yes/no	e.g. figshare DOI or NCBI bioproject accession number	yes/no	[insert any comments about the data and the process finding it]	

Colour code: can be assigned to final results to assist with scoring:
Green---Data available (by standards of the field)
Orange---Flagged ethical issues with data access/need to request access
Red---Data not available

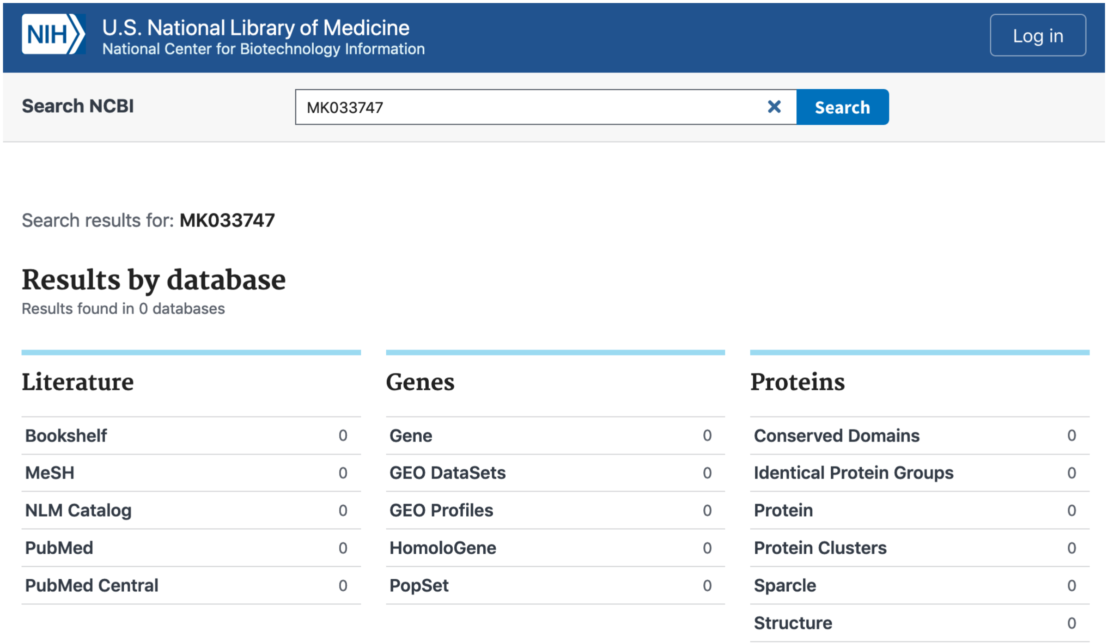
3 Students check the papers and summarise results.

This step is provided as a separate protocol for students, see:dx.doi.org/10.17504/protocols.io.6yahfse

4 Assessment of crowdsourced results

Once the students have completed the assignment, work through and mark the results spreadsheet. If necessary providing a quick double check of the papers to see if the students were correct in their overall assessment of the data availability.

4.1 Check if external data sources (data DOIs or accessions) work, and make a note when marking the students if you agree with their assessments. If the students have noted that "data is available on request", the availability of this data can be checked by sending an email to the contact listed on the paper.


NIH
U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
MK033747
Search

Search results for: **MK033747**

Results by database

Results found in 0 databases

Literature	
Bookshelf	0
MeSH	0
NLM Catalog	0
PubMed	0
PubMed Central	0

Genes	
Gene	0
GEO DataSets	0
GEO Profiles	0
HomoloGene	0
PopSet	0

Proteins	
Conserved Domains	0
Identical Protein Groups	0
Protein	0
Protein Clusters	0
Sparcle	0
Structure	0

Checking an external data source: in this case testing if a GeneBank accession number is currently publicly available in the NCBI databases.

5 Assessment of "data available on request"

5.1 If the paper assessed says "data available on request" this can be tested by sending a simple request to the highlighted contact or data access committee. Test if the email address works and if they reply by sending a simple email to the listed contact like this one:

Dear [insert name],

Is it possible to get hold of the [de-identified/anonymised] data covered in this publication:
[insert paper DOI here]
The paper states data are available to interested researchers upon request.

Sincerely,

If the paper states the supporting data is of a clinical/medical nature, always request de-identified/anonymised data.

It is not necessary to get hold of the data, but just note if the email address works and if the corresponding author/data access contact responds. If there is no response within a week make a note of this and email again. If there is still no response then mark the data in the spreadsheet as missing.

6 Calculate final results

Using the results in the spreadsheet you can now calculate what proportion of the data is available without restrict, is available on request, and is missing. You can also assess the students on how accurate their data is compared to your assessment. If you want to share the final results you may want to make a corrected copy and remove the student details to protect their identity. For an example of these outputs you can the results in figshare from our previous student project results.