

“We are the Champions”: Exploring a Data Champions Pilot in the Canadian Context

**Jane Fry, Nick Rochlin,
Colin Conrad, Jen Pecoskie**

June 02, 2023 // 02 juin 2023



Digital Research
Alliance of Canada

Alliance de recherche
numérique du Canada



Agenda

1. Introduction
2. Introduction to Data Champions
 - a. Jen Pecoskie
3. Individual Data Champions Projects
 - a. Jane Fry
 - b. Colin Conrad
 - c. Nick Rochlin
4. Panel Discussion with Audience Input

The Alliance's Data Champions Pilot

<https://alliancecan.ca/en/funding-opportunities/data-champions-pilot-project-call>

About ▾ Membership ▾ Services ▾ Funding Opportunities ▾ Initiatives ▾ Latest ▾

Contact

About

Membership

Services

Funding
Opportunities

Initiatives

Latest

Contact

Data Champions Pilot Project Call

Data Champions Pilot Project Call is now closed.

The Alliance is pleased to announce its Data Champions Pilot Project Call (\$916,000 CAD). This funding call aims to build national research capacity and deliver on the [Alliance's mandate](#) to create a broad and integrated Canadian digital research infrastructure (DRI) ecosystem. Specifically, the call will address the needs of the research community related to Research Data Management (RDM), while promoting an equitable and inclusive DRI environment in Canada.

Important information:

Broad Areas of Data Champion Activity



Data Champions - Funding Call

- ▶ Specific eligibility criteria
 - ▶ University, post-secondary college, educational institution, or hospital
 - ▶ Non-profit organization
- ▶ Duration - 1 year, from April 2022 - March 2023
- ▶ Awards up to \$50,000 CDN
 - ▶ Salaries / stipends and pay-related benefits
- ▶ Merit-review process

Data Champions - Awardees and Support

- ▶ 18 Data Champions named
 - ▶ From across Canada, in English and French, representative of institutional size and type
 - ▶ <https://alliancecan.ca/en/funding-opportunities/data-champions-pilot-project-call>
- ▶ Community of Practice over the award year
 - ▶ Supported by an Organizing Committees and external stakeholders

Data Champions - Successful?

1. A cohort of Data Champions (composed of individuals and/or teams) individually and/or collaboratively develops, promotes, contributes to, and/or delivers on a range of RDM activities over the course of the Pilot.
2. Data Champion activities increase awareness of and competence in RDM methods and practices within disciplines, regions, and/or institutions, as measured by means identified in the applications.
3. A National Community of Practice is established, relationships are built, and information is gathered to inform parameters for a potential future Data Champions Program.

Measure of Success #2:

“there is now much more awareness of RDM and DMPs on campus. Researchers have begun asking more questions and many are soliciting advice.”

“[in] advancing practice at [our University name] and nationally, ... we believe we have begun to do that, and have built the capacity and insight to do far, far more.”

Data Champions - Successful?

1. A cohort of Data Champions (composed of individuals and/or teams) individually and/or collaboratively develops, promotes, contributes to, and/or delivers on a range of RDM activities over the course of the Pilot.
2. Data Champion activities increase awareness of and competence in RDM methods and practices within disciplines, regions, and/or institutions, as measured by means identified in the applications.
3. A National Community of Practice is established, relationships are built, and information is gathered to inform parameters for a potential future Data Champions Program.

Measure of Success #3:

“The greatest benefit of being part of this community was the possibility of learning through the exchange of experiences and knowledge. We hope this community remains active and open for future collaborations. Our goals in advancing RDM and implementing best practices across Canada overlap. Therefore, developing projects in collaboration with the community would be great.”



Thank you!

Jen Pecoskie

jen.pecoskie@alliancecan.ca

01

02

03

04

05

06

07

08

Over to Jane Fry!

Thanks to ...

- The Digital Research Alliance of Canada (the Alliance)
 - For the funding that allowed us to undertake this work.
- Jen Pecoskie (RDM Project Coordinator, the Alliance)
 - For our regular meetings, keeping us on track and answering all questions, no matter how trivial!

At Carleton - Thanks to ...

- The Carleton Office for Research Initiatives and Services (CORIS)
 - And Andrea Lawrance (Director of CORIS)
- The three graduate Research Assistants:
 - Anamika Jayendran; Stephan Struve; and Amara Umeh

Our Project

- Title:
 - *Beyond the RDM Checkbox: An RDM Webinar Series to build expertise for all*
- Carleton U Working Group
 - Andrea Lawrance, Director, CORIS
 - PI: Jane Fry, Data Services Librarian, Carleton U
 - 3 grad RAs
 - Various other advisors
- Deliverables
 - On the Carleton RDM webpage
 - <https://library.carleton.ca/rdm-learning-resources-faq>

Deliverable - Webinar Series

- *Introduction to RDM*
 - *Data Storage*
 - *Data Management Plans*
 - *The Nitty Gritty of RDM: Part 1 - Supporting Researchers*
 - *The Nitty Gritty of RDM: Part 2 - The Researchers' Experience*
-
- All took place over Zoom, from January to March 2023
 - 90 minutes in length
 - Open to Carleton community and beyond
 - Conducted a post-webinar survey

Webinar Series (cont'd)

- I was the organizer and moderator
- Recruited experts in the field to be presenters
- Recorded webinars
- Participants
 - From over 70 different institutions
 - Universities
 - Colleges
 - Research institutions at hospitals
 - Various gov't agencies
 - From across Canada
 - and Internationally
 - Over 600 registrants
 - Most popular one - DMPs

Deliverable - Environmental Scan

- On RDM Training Materials
- Done by 3 RAs
 - This helped to indoctrinate them into RDM
- They were also to look for gaps in training
 - This would help them determine what Learning Modules needed to be created
- There is a Version 1 ready
 - I plan to add to it, with a historical perspective and additional materials

Deliverable - Learning Modules

- Done by RAs
 - Interactive (mostly)
 - Self-directed
-
- *Introduction to RDM*
 - *Data Management Plans*
 - *Data Sharing*
 - *Institutional Learning*
 - *Organizing Data*

Other Deliverables

- Done by RAs
- *RDM FAQs*
 - *From questions asked at the webinars*
- *RDM Training Materials Repository*
 - *Created from their Environmental Scan*

Benefits for me

- Updating the Introductory training materials that I had done over 5 years ago
 - As Chair of the Portage Training Expert Group (a national group)
 - This makes it much easier for me when I am teaching and doing consultations
- Indoctrinating 3 grad students into RDM!

Challenges

- Importance of being organized throughout the project
- We are a busy group of people!
- The HR stuff

Thank you!

For more information ...

<https://library.carleton.ca/rdm-learning-resources-faq>

Jane Fry

jane.fry@carleton.ca

01

02

03

04

05

06

07

08

Over to Colin Conrad!

At Dalhousie... funding and thanks

Thanks to ...

- Dr. Darren Abramson for helping conceptualize this.
- Juan Chaves Baquero, Maddie Hare and Poppy Riddle who did most of the work through their research.
- The Digital Research Alliance of Canada for supporting these three to do the work!

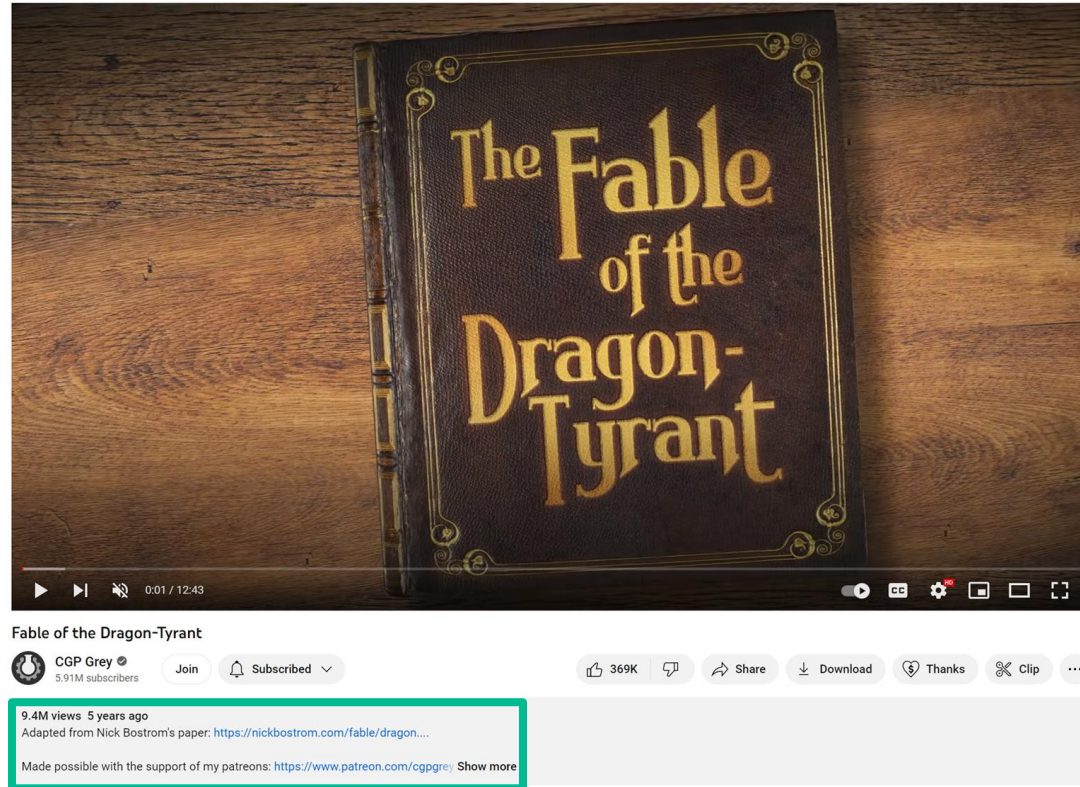


Digital Research
Alliance of Canada

Alliance de recherche
numérique du Canada

The Inspiration

- My former MA supervisor knew the essay, but not the video.
- “It’s not **important** for my work.”
- If YouTube is not important for philosophy, what other gaps are there?



CGP Grey (Apr 24, 2018). Fable of the Dragon-Tyrant. YouTube. <https://www.youtube.com/watch?v=cZYNADOHhVY&t=94s>

The challenge

1. Is there a gap in the digital and data skills among social sciences and humanities researchers?
2. Is there an appetite for learning digital and data skills?
3. What are the barriers to the adoption of digital tools and data oriented research?

Our approach

Computational social sciences research resource (Juan)	Digital humanities teaching resource (Poppy)	Research project about perceived barriers (Maddie).
Demonstrate open access research related to digital technology and social science. Topic: Adoption of GDPR compliant privacy notifications (paper forthcoming).	Prepare a workshop to equip DH scholars with digital skills. Provide supporting digital content. Result: https://digitalhumanities.github.io/DH_Topic_Workshop	Through interviews and surveys identify perceived barriers to the adoption of digital humanities tools. Result: Qualitative factors (forthcoming).

Manateam Documentation on OCR and LDA



Welcome to our documentation introducing you to Python-based digital humanities tools.

What is Digital Humanities?

Digital Humanities (DH) is the application of digital tools to process information to enable researchers to explore subjects in the humanities in new ways. Digital tools can provide advantages to the researcher with processing capabilities on huge amounts of information on texts, images, sound, or even on data itself. While DH employs digital technology with the goal is deriving new insights, these same tools can also be the subject of critical inquiry as well. It is a huge field with simultaneous developments in other fields. The [Wikipedia entry](#) on DH is well worth reading for an introduction, as are the resources from [University of Victoria](#), [Stanford](#), and the [University of California Berkeley](#).

Our workshop will focus on text analysis in which we look for topics, though there are many types of text analysis.

This documentation covers the following:

- Chapter 1 - Getting started
- Chapter 2 - Finding source material
- Chapter 3 - Using the OCR notebook
- Chapter 4 - Using the LDA notebook
- Chapter 5 - Moving on and using your own data

Objectives

Our workshop seeks to provide the following for our participants:

- Use Python to run an Optical Character Recognition (OCR) library called PyTesseract to extract text from a scanned PDF file and perform topic modeling using LDA to derive meaningful topics.
- Create visualizations, with both Python and Tableau Public, to generate meaningful topics from the text.

On this page

[What is Digital Humanities?](#)

This documentation covers the following:

[Objectives](#)

[What is OCR?](#)

[What is LDA?](#)



Source material

AUTHOR
the Manateam

On this page
Finding source material
Other source materials
What it can't do

Finding source material

PDF files work best for this analysis, but there are three types of PDF files to be aware of:

- digitally-created pdfs such as from print streams in Word, that contain text and font-family information, and are searchable.
- scanned/image-only pdfs that contain no text information
- searchable pdfs - where there is a text layer that has been developed with OCR and sits underneath the scanned image

Archives will typically use scanned image-only pdfs, though sometimes these are saved as PDF/A, which is a reduced file size with reduced functionality. For our Gazette project, you can find high quality scans of documents [here at the Dalhousie Gazette Archives](#). You can find the current Gazette [here](#).



The following is an example screenshot of a multi-page scanned image PDF file of a 1922 Dalhousie Gazette.



save it somewhere, like GitHub or locally on your computer.

Process guide for the OCR notebook

Step 1:

There are two ways to open the OCR notebook:

1. open through Github [insert screenshot] - you will use this process the first time opening the notebooks.

1. Save a copy of this notebook to your Google Drive. All changes will be saved to your own copy.

2. Open through Colab [insert screenshot] - You can use this process if you have previously saved it to your Google Drive.

1. In Colab, go to File>Open. This will open a dialog box.

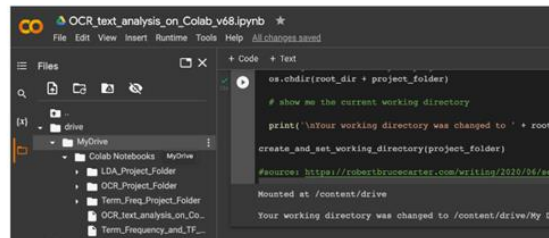
2. At the top tabs of the dialog box, select Google Drive - you should see your previously saved copy.

Run the first cell - this will connect to your Google Drive. A dialog box will open and ask you to select your Google account and ask for your permission to connect to your Google Drive.

It will also create a new directory in your Google Drive for all of the files needed for this project. You should see something like this indicating your new working directory.

```
Mounted at /content/drive
Your working directory was changed to /content/drive/My Drive/Colab Notebooks/OCR_Project_Folder/
```

Follow the steps for 1d to put the scanned pdf files into the working directory. Colab has a file manager on the left hand side that will allow you to drag and drop files to the working directory, OCR_Project_Folder.



```
OCR_text_analysis_on_Colab_v68.ipynb *
File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

os.chdir(root_dir + project_folder)
# show me the current working directory
print('!Your working directory was changed to ' + root_
create_and_set_working_directory(project_folder)
#source: https://robertbrucearter.com/writing/2020/06/set-
Mounted at /content/drive
Your working directory was changed to /content/drive/My Dr
```

On this page

Using the OCR notebook on Colab

Objectives

About Google Colab

Process guide for the OCR notebook

Step 1:

Step 2:

Step 3:

Step 4:

Managing output from the notebook

Future improvements & applications



Key findings

1. Many researchers are like my former supervisor: “this is not important for what I do.” (low perceived usefulness)
2. Perceived challenges with power structures, financial interests, and computer science disciplines.
3. Differences in what is meant by “digital humanities”.
4. Interest in the hands-on teaching approach.

Thank you!

Colin Conrad, PhD

Colin.Conrad@Dal.Ca



Manateam website:

https://digitalhugmanitees.github.io/DH_Topic_Workshop

01

02

03

04

05

06

07

08

Over to Nick Rochlin!

Funding

- ▶ Special thanks to Jen Pecoskie and Laura Gerlitz!



**Digital Research
Alliance** of Canada

**Alliance de recherche
numérique** du Canada

The Team

Admin Units

- ▶ **Library**
 - ▶ Marjorie Mitchell - RDM Librarian
 - ▶ Mathew Vis-Dunbar - Data & Digital Scholarship Librarian
- ▶ **ARC**
 - ▶ Nick Rochlin, RDM Specialist

Research Labs

- ▶ **Health & Exercise Science / Integrated Knowledge Translation** - Dr. Heather Gainforth
 - ▶ Kelsey Wuerstl
 - ▶ Kailan Tang
- ▶ **Visual Anthropology** - Dr. Fiona McDonald
 - ▶ Hanna Paul
 - ▶ Morgan King
- ▶ **English & Cultural Studies / Digital Humanities** - Dr. Emily Murphy
 - ▶ Craig Jacobs
- ▶ **Biology / Conservation and Ecology** - Dr. Jason Pither
 - ▶ Liam Johnson
 - ▶ Jordan Katchen

The issue

- ▶ There's a lot of RDM materials available, but...
 - ▶ Generic resources can be difficult to translate to practical, discipline-specific applications
 - ▶ Most resources are geared at the project level, which misses a lot of RDM troubles

An approach to a solution

- ▶ Create RDM resources geared to the research lab
- ▶ Facilitate the creation of RDM sections in lab manuals
- ▶ Focus on grad student experience

Proposed Project Deliverables

1. A high-level RDM resource for graduate labs
2. Customizable templates for RDM lab workshops
3. Student-led implementation of lab workshops
4. Exemplar implementations of RDM manuals in each lab
5. A knowledge exchange network

What actually happened...



What actually happened...

1. A high-level RDM resource for graduate labs

What actually happened...

1. A high-level RDM resource for graduate labs
2. **Customizable templates for RDM lab workshops**

What actually happened...

1. A high-level RDM resource for graduate labs
2. Customizable templates for RDM lab workshops
3. **Student-led implementation of lab workshops**

What actually happened...

1. A high-level RDM resource for graduate labs
2. Customizable templates for RDM lab workshops
3. Student-led implementation of lab workshops
4. **Exemplar implementations of RDM manuals in each lab**

What actually happened...

1. A high-level RDM resource for graduate labs
2. Customizable templates for RDM lab workshops
3. Student-led implementation of lab workshops
4. Exemplar implementations of RDM manuals in each lab
5. **A knowledge exchange network**



Thank you!

Nick Rochlin

nick.rochlin@ubc.ca

01

02

03

04

05

06

07

08

IASSIST Panel Members

- Jen Pecoskie
 - RDM Project Coordinator, The Alliance
 - jen.pecoskie@alliancecan.ca
- Jane Fry
 - Data Services Librarian, MacOdrum Library, Carleton University
 - jane.fry@carleton.ca
- Colin Conrad
 - Assistant Professor, Faculty of Management, Dalhousie University
 - Colin.Conrad@Dal.Ca
- Nick Rochlin
 - RDM Specialist, University of British Columbia
 - nick.rochlin@ubc.ca



Digital Research Alliance of Canada

Accelerating Canada's
Research Future.

Alliance de recherche numérique du Canada

Accélérer l'avenir de
la recherche au Canada.

Thank you!
alliancecan.ca

