

# TurboCurator for Metadata Curation in Dataverse

An update at Scholars Portal Day  
May 1, 2024

Amber Leahey, Data & GIS Librarian | Borealis Service Director (Scholars Portal)

# LLM AI background & data curation

- Explosion of AI applications across scientific projects, spanning all forms of academic work to improve outputs and advance scientific discovery
- For research data, this includes text and data cleaning, analysis, reuse and now metadata curation
- A built-in repository solution makes this really easy to get started
- LLM AI and reuse of data raises important questions:
  - Data licensing (e.g. CC0)
  - LLM AI's accuracy and precision
  - Openness and flexibility of tools to apply to different use cases and customize underlying training & learning models (e.g. best practices)
  - What is the user experience? Predictability and often a lot of unknowns...
- Data repositories 'use cases'
  - Data science & scientific discovery
  - Data Curation approaches
    - Automated curation and applying curation best practices and standards
  - And more...

# ICPSR TurboCurator

- ICPSR is a world renown social science data archive
  - Provides archiving of U.S. and international research data including surveys, political opinion data, government data and a variety of research data from across collaborative initiatives
- Dataverse community & TurboCurator
  - Dataverse is a repository for depositing and publishing research data
  - Widespread open source community of repository software adopters with invariable curation needs
  - TurboCurator is built into Dataverse and uses ICPSR's curation best practices and Open AI's ChatGPT
    - Metadata integration supports suggested AI improvements for Title, Abstract (Description), and Keywords present in the Dataverse Dataset metadata fields
    - Can be used as part of deposit process or as a tool to improve existing metadata in collections in Dataverse
    - **\*Does not support data files, only metadata (intended to support files in the future)**
    - **\*Only use with non-sensitive data**

# TurboCurator

**Info** – This draft version needs to be published. When ready for sharing, please **publish** it so that others can see these changes.

## Testing Curation Features

Draft Unpublished



Leahey, Amber, 2024, "Testing Curation Features", <https://doi.org/10.70122/FK2/L7CNP1>, Demo Dataverse, DRAFT VERSION, UNF:6:69/RG24IZRqkomroYSG5iA== [fileUNF]

Cite Dataset

Learn about [Data Citation Standards](#).

### Description

Test

### Subject

Arts and Humanities; Astronomy and Astrophysics; Business and Management; Chemistry; Earth and Environmental Sciences; Engineering; Medicine, Health and Life Sciences; Computer and Information Science; Law; Mathematical Sciences; Physics; Social Sciences; Other; Agricultural Sciences; Demo Only

### Keyword

test

### Related Publication

test

### Notes

test

### License/Data Use Agreement



CC0 1.0

Files

Metadata

Terms

Versions

+ Upload Files

1 File

Edit Files



LFS\_January\_2024.tab

Tabular Data - 10.8 MB

- Access Dataset
- Publish Dataset
- Edit Dataset
- Contact
- Dataset Metadata
- 0 Downloads
- Files (Upload)
- Metadata
- Terms
- Permissions
- Private URL
- Thumbnails + Widgets
- Configure Options
- TurboCurator by ICPSR
- Delete Dataset

# Launch TurboCurator

Top navigation: Add Data, Search, About, User Guide, Support, Amber Leahey 16

Page title: Demo Dataverse >

Info banner: Info – This draft version needs to be published. When ready for sharing, please publish it so that others can see these changes.

## Testing Curation Features

Draft Unpublished

Leahey, Amber, 2024, "Testing Curation Features", <https://doi.org/10.70122/FK2/L7CNP1>, Demo Dataverse, DRAFT VERSION, UNF:6:69/RG24IZRqkomroYSG5iA== [fileUNF]

Cite Dataset Learn about [Data Citation Standards](#).

Buttons: Access Dataset, Publish Dataset, Edit Dataset, Contact Owner, Share

Dataset Metrics: 0 Downloads

### TurboCurator by ICPSR

TurboCurator is a configuration tool as it wants to suggest changes prior to deposit and eventually allow users to push changes back to their dataset metadata. The longer term plan is also to make suggestions to files.

Continue Cancel

Description: Test

Subject: Arts and Environment Science Only

Keyword: test

Related Publication: test

Notes: test

License/Data Use Agreement: PUBLIC DOMAIN CC0 1.0

Navigation: Files Metadata Terms Versions


Buttons: + Upload Files, Edit Files

File list: 1 File

- LFS\_January\_2024.tab  
Tabular Data - 10.8 MB

## Tell us more about your deposit

[Why are we requesting more information?](#)

 Do not enter any sensitive information that may contain direct or indirect identifiers in your data. Information entered here will be sent to Open AI's ChatGPT. [ICPSR's data confidentiality policy \(new tab\)](#).

### Enter additional information that might offer improved suggestions

We encourage you to add any additional information you have about your data or research into the box below.

### What kind of information are we looking for?

#### Abstract or summary about your research

This could come from a published paper or a press release on your study. If you don't have an abstract or summary on hand, give us a two-sentence description of your study.

#### Your data collection and analysis methodology

This could be the methodology section of your published paper or research plan.

#### Deposit treatment information

Did you take any extra or important steps, such as a special analysis method? Were there data that you collected in the research that were not uploaded?

By default takes input from DV:

- Title
- Description
- Keywords

Any additional information not included can be added to form

Additional input types

Do not enter any sensitive information that may contain direct or indirect identifiers in your data. Information entered here will be sent to Open AI's ChatGPT. [ICPSR's data confidentiality policy \(new tab\)](#).

**Enter additional information that might offer improved suggestions**

We encourage you to add any additional information you have about your data or research into the box below.

This is the monthly release of the Labour Force Survey for January 2024. Survey responses are from Canada.

**What kind of information are we looking for?**

**Abstract or summary about your research**

This could come from a published paper or a press release on your study. If you don't have an abstract or summary on hand, give us a two-sentence description of your study.

**Your data collection and analysis methodology**

This could be the methodology section of your published paper or research plan.

**Deposit treatment information**

Did you take any extra or important steps, such as a special analysis method? Were there data that you collected in the research that were not uploaded?

**Miscellaneous Information**

Is there any additional information about your study or data that was not included in the deposit form?

Hide

- By checking this box, I acknowledge and accept that the metadata and information I entered will be shared with OpenAI's ChatGPT. [TurboCurator's data sharing agreement \(new tab\)](#).
- By checking this box, I acknowledge and accept that TurboCurator by ICPSR uses and logs non-sensitive metadata. [How does TurboCurator work? \(new tab\)](#)

Submit & Show Recommendations

Do Not Generate Recommendations

Helpful Documents:

- [About TurboCurator](#) (ICPSR)
- [FAQ](#)

Terms of Use (\*only use w/ non-sensitive data\*)

'Submit' for AI suggestions

# Improved 'Title'

## Title

[What are these suggestions based on?](#)

Labour Force Survey: January 2024 [Canada]

 Copy Title Recommendation

Your Entry



Testing Curation Features

 Hide

---

## Recommendation Rules: Title




Title recommendations are based on ICPSR's rules and best practices for depositing data.

### Geography

- Geographical location must be added if not present;
- City and state should be added to the title, if applicable;
- If "National," "American," "Congressional," or "Federal" is in the title, add [United States] in brackets at the end of the title;
- Geographic location should not be written in all caps, but be written as such: United States; Philadelphia County, Pennsylvania;
- Spell out the state name of United States and Canada;
- If there are more than 4 cities, countries, regions, etc. in the Title, then specify geographic information as "10 Caribbean countries" or "5 European cities";

### Time Period

- All titles should reflect the time period that the study covers.

 If you have any questions with respect to these rules, please email us [ICPSR-help@umich.edu](mailto:ICPSR-help@umich.edu)

Close

# No 'Description' suggested

## Description

[What are these suggestions based on?](#)

The summary is not provided. Please provide the summary to revise.

 Copy Description Recommendation

Your Entry



## Recommendation Rules: Description



Description recommendations are based on ICPSR's rules and best practices for depositing data.

- It is important to indicate the unit of analysis for the data collection, i.e., who or what is being studied: individuals, housing units, courts, criminal acts, etc. Most often the unit of analysis is the individual; where it is not, it is particularly important to make this clear.
- Always use numbers versus spelling it out. Spelling out the number can sometimes be used for emphasis, but in that case, the number should also be used in parenthesis (e.g. Two thousand (2,000)).
- Use past tense when describing the process of collecting the data. Use present tense when necessary, such as when you are describing the data: "The MIDUS Refresher collection is split into two datasets.




If you have any questions with respect to these rules, please email us [ICPSR-help@umich.edu](mailto:ICPSR-help@umich.edu)

Close

# No 'Keywords' suggested

## Keywords

[What are these suggestions based on?](#)

 Sorry, we were unable to recommend any keywords that match the list of [ICPSR preferred keywords](#).

Your Entry



## Recommendation Rules: Keywords




Keywords recommendations are based on ICPSR's rules and best practices for depositing data.

- Keywords suggestions are based on the preferred keywords list from the [ICPSR Thesaurus](#).




If you have any questions with respect to these rules, please email us [ICPSR-help@umich.edu](mailto:ICPSR-help@umich.edu)

Close

-  Important Note:
- The recommendations below will not be reflected in your deposit.
  - To use the recommended text, use the copy button and then paste the updated text into your deposit.

# Test 2: Add more input metadata

- Old description from December 2023 LFS Study
- Old keywords from December 2023 LFS


 Add Data ▾ Search ▾ About User Guide Support Amber Leahey 15

Demo DataVerse >

**!** Info – This draft version needs to be published. When ready for sharing, please **publish** it so that others can see these changes.

## Labour Force Survey: January 2024 [Canada]

**Draft** **Unpublished**

 Leahey, Amber, 2024, "Labour Force Survey: January 2024 [Canada]", <https://doi.org/10.70122/FK2/L7CNP1>, Demo DataVerse, DRAFT VERSION, UNF:6:69/RG24IZRqkomroYSG5IA== [fileUNF] **?**

[Cite Dataset ▾](#)    Learn about [Data Citation Standards](#).

**Description** **?**    [old description - to be updated:] The Labour Force Survey (LFS) is a monthly survey of Canadian households carried out by Statistics Canada. It was developed after the Second World War to satisfy a need for reliable and timely data on the labour market due to the massive labour market changes involved in the transition from a war to peace-time economy. The objectives of the LFS have been to divide the working-age population into three mutually exclusive labour force status categories (employed, unemployed, and not in the labour force) and to provide descriptive and explanatory data on each of these groups. With the release of the survey results only 10 days after the completion of data collection, the LFS estimates are the first of the major monthly economic data series to be released. The LFS is the source of Canada's official unemployment rates, including the rates used by Employment and Social Development Canada in the estimation of Employment Insurance (EI).


[Read full Description \[+\]](#)

**Subject** **?**    Arts and Humanities; Astronomy and Astrophysics; Business and Management; Chemistry; Earth and Environmental Sciences; Engineering; Medicine, Health and Life Sciences; Computer and Information Science; Law; Mathematical Sciences; Physics; Social Sciences; Other; Agricultural Sciences; Demo Only

**Keyword** **?**    labour force, occupation, income, job

**Related Publication** **?**    test

**Notes** **?**    test

**License/Data Use Agreement**     **PUBLIC DOMAIN** CC0 1.0

[Files](#)   [Metadata](#)   [Terms](#)   [Versions](#)

[Access Dataset ▾](#)

[Publish Dataset ▾](#)

[Edit Dataset ▾](#)

[Contact Owner](#)    [Share](#)

**Dataset Metrics** **?**

0 Downloads **?**

[+ Upload Files](#)

## Test 2: Improved Description

### Description

[What are these suggestions based on?](#)

The Labour Force Survey (LFS) is a monthly survey of Canadian households carried out by Statistics Canada, serving as the official source of unemployment rates. The survey aims to provide reliable and timely data on the labour market by categorizing the working-age population into employed, unemployed, and not in the labour

 Copy Description Recommendation

Your Entry

[old description - to be updated:] The Labour Force Survey (LFS) is a monthly survey of Canadian households carried out by Statistics Canada. It was developed after the Second World War to satisfy a need for reliable and timely data on the labour market due to the massive labour market changes involved in the transition from a war to peace-time economy. The objectives of the LFS have been to divide the working-age population into three mutually exclusive labour force status categories (employed, unemployed, and not in the labour force) and to provide descriptive and explanatory data on each of these groups. With the release of the survey results only 10 days after the completion of data collection, the LFS estimates are the first of the major monthly economic data series to be released. The LFS is the source of Canada's official unemployment rates, including the rates used by Employment and Social Development Canada in the calculation of Employment Insurance (EI) eligibility and benefit criteria. Data from the survey also provide information on major labour market trends, such as shifts in employment across industrial sectors, hours worked and labour force participation. The LFS also provides employment estimates by industry, occupation, public and private sector, hours worked and much more, all cross-classifiable by a variety of demographic characteristics. Estimates are produced for Canada, the provinces, the territories and a large number of sub-provincial regions. For employees, data on wage rates, union status, job permanency and establishment size are also produced.

# Original Input ‘Description’

“[old description - to be updated:] The Labour Force Survey (LFS) is a monthly survey of Canadian households carried out by Statistics Canada. It was developed after the Second World War to satisfy a need for reliable and timely data on the labour market due to the massive labour market changes involved in the transition from a war to peace-time economy. The objectives of the LFS have been to divide the working-age population into three mutually exclusive labour force status categories (employed, unemployed, and not in the labour force) and to provide descriptive and explanatory data on each of these groups. With the release of the survey results only 10 days after the completion of data collection, the LFS estimates are the first of the major monthly economic data series to be released. The LFS is the source of Canada's official unemployment rates, including the rates used by Employment and Social Development Canada in the calculation of Employment Insurance (EI) eligibility and benefit criteria. Data from the survey also provide information on major labour market trends, such as shifts in employment across industrial sectors, hours worked and labour force participation. The LFS also provides employment estimates by industry, occupation, public and private sector, hours worked and much more, all cross-classifiable by a variety of demographic characteristics. Estimates are produced for Canada, the provinces, the territories and a large number of sub-provincial regions. For employees, data on wage rates, union status, job permanency and establishment size are also produced.”

## Improved AI generated 'Description'

“The Labour Force Survey (LFS) is a monthly survey of Canadian households carried out by Statistics Canada, serving as the official source of unemployment rates. The survey aims to provide reliable and timely data on the labour market by categorizing the working-age population into employed, unemployed, and not in the labour force, and by providing descriptive and explanatory data on each group. The LFS releases survey results only 10 days after data collection, making it the first major monthly economic data series to be released. It provides information on major labour market trends, such as shifts in employment across industrial sectors, hours worked, labour force participation, employment estimates by industry, occupation, public and private sector, and more. The survey also produces data on wage rates, union status, job permanency, establishment size, and is based on responses from Canadian households.”

# Improved AI generated Description - Difference check

“The Labour Force Survey (LFS) is a monthly survey of Canadian households carried out by Statistics Canada, serving as the official source of unemployment rates. The survey aims to provide reliable and timely data on the labour market by categorizing the working-age population into employed, unemployed, and not in the labour force, and by providing descriptive and explanatory data on each group. The LFS releases survey results only 10 days after data collection, making it the first major monthly economic data series to be released. It provides information on major labour market trends, such as shifts in employment across industrial sectors, hours worked, labour force participation, employment estimates by industry, occupation, public and private sector, and more. The survey also produces data on wage rates, union status, job permanency, establishment size, and is based on responses from Canadian households.”

The differences between the two blocks of text are:

- The AI suggested description is shorter length (139 words) than the original (242 words).
  - In the original block, it is mentioned that the LFS was developed after the Second World War to address the changes in the labour market from war to peace-time economy, while the right block does not include this historical context.
  - The original block specifies that the LFS provides data cross-classifiable by a variety of demographic characteristics, while the right block does not mention this.
  - The original block mentions that the LFS provides data for Canada, provinces, territories, and sub-provincial regions, while the AI generated block does not include specific details on geographical coverage.
  - The original block includes information about data on Employment Insurance eligibility and benefit criteria, while the right block does not mention this.
  - The original block emphasizes the production of data on employment estimates by industry, occupation, public and private sector, hours worked, and much more, while the AI generated block provides a more general overview.

# Found 1 matching 'Keyword'

## Keywords

[What are these suggestions based on?](#)

income

 Copy Keywords Recommendation

Your Entry



## Data deposit recommendations for Labour Force Survey: January 2024 [Canada]

Date Generated: 4/29/2024

# Seems to improve every time you run it...

Ran this same draft dataset again  
(same input) in TurboCurator on **April  
29, 2024**

- Improvements to geographic coverage in 'Description'
- More keywords matched/recognized

- Review your recommendations below and update as needed
- Copy and paste in your deposit to update it

### Title

[What are these suggestions based on?](#)

Labour Force Survey: January 2024 [Canada]

[Copy Title Recommendation](#)

Your Entry

### Description

[What are these suggestions based on?](#)

sector, hours worked, and other demographic characteristics. Wage rates, union status, job permanency, and establishment size data are produced for employees. The data are available for Canada, provinces, territories, and sub-provincial regions.

[Copy Description Recommendation](#)

Your Entry

### Keywords

[What are these suggestions based on?](#)

employment, statistics, unemployment, income

[Copy Keywords Recommendation](#)

Your Entry

# Uses Cases & Applications for Borealis

## Depositors & Curators

- Training students and researchers in metadata creation & data cataloguing
- New deposits review w/ TurboCurator
  - Research data are submitted for review by researchers to institutional admins in Borealis
  - Depositors/Admins/Curators review metadata using TurboCurator to suggest improvements to authors/depositors before publishing
- QA/QC with suggested improvements in collections
  - Curated and existing collections have metadata that can be improved for consistency and best practices (e.g. Odesi, institutional collections, etc.)
  - Improve Titles, Description, and Keywords based on best practices
    - Odesi Metadata Best Practices
    - Odesi Deposit Guide
    - Ex. UVic Deposit Guidelines
- Remains unclear whether we can use our own best practices (or just ICPSR's)

# Broader Dataverse & AI possibilities...

- Multilingual search and cataloguing (Enter French search, get French and English only results)
- Semantic search (e.g. income = salary without having to use these terms)
- Data Analysis ('Ask the Data' tool analyses data files e.g. "What is the relationship between income and education?")
- Large data computation without having to 'move' or 'share' datasets
- Cataloguing and metadata curation (e.g. TurboCurator)
- Other possibilities...
  - Data preservation
  - Data Science (e.g. data and variable interoperability)

Demo Dataverse >

## TEST Labour Force Survey: January 2024 [Canada]

Version 1.0

Leahy, Amber, 2024, "TEST Labour Force Survey: January 2024 [Canada]", <https://doi.org/10.70122/FK2/L7CNP1>, Demo Dataverse, V1, UNF:6.69/RG24ZRqkomoYSG5A== [fileUNF]

[Cite Dataset](#) - [Learn about Data Citation Standards](#)

[Access Dataset](#) -  
[Edit Dataset](#) -  
[Link Dataset](#)  
[Contact Owner](#) [Share](#)

[Dataset Metrics](#) ⓘ  
0 Downloads ⓘ

**Description** ⓘ

[old description - to be updated] The Labour Force Survey (LFS) is a monthly survey of Canadian households carried out by Statistics Canada. It was developed after the Second World War to satisfy a need for reliable and timely data on the labour market due to the massive labour market changes involved in the transition from a war to peace-time economy. The objectives of the LFS have been to divide the working-age population into three mutually exclusive labour force status categories (employed, unemployed, and not in the labour force) and to provide descriptive and explanatory data on each of these groups. With the release of the survey results only 10 days after the completion of data collection, the LFS estimates are the first of the major monthly economic data series to be released. The LFS is the source of Canada's official unemployment rates, including the rates used by the International Labour Organization in its publications on unemployment statistics.

[Read full Description](#) [+]

**Subject** ⓘ Arts and Humanities; Astronomy and Astrophysics; Business and Management; Chemistry; Earth and Environmental Sciences; Engineering, Medicine, Health and Life Sciences; Computer and Information Science; Law, Mathematical Sciences, Physics; Social Sciences; Other, Agricultural Sciences; Demo Only

**Keyword** ⓘ labour force, occupation, income, job

**Related Publication** ⓘ test


**Notes** ⓘ test

**License/Data Use Agreement**  CC BY 4.0



[Files](#) [Metadata](#) [Terms](#) [Versions](#)

[+ Upload Files](#)

**1 File** [Edit Files](#) -

 LFS\_January\_2024.tab  
Tabular Data - 10.8 MB  
Published Apr 29, 2024  
0 Downloads  
60 Variables · 105278 Observations UNF:6.69/R...5/A== ⓘ  
LFS Jan 2024  
[Data](#) [Jan 2024](#)

**Ask the Data**

# 'Ask the Data' tool

Already reads file level  
metadata (e.g. .Tab of  
CSV, Excel, SPSS,  
STATA)

"Give me all the variable  
labels"

Version 1.0

File Citation

Leahey, Amber, 2024, "LFS\_January\_2024.tab", TEST Labour Force Survey: January 2024 [Canada], <https://doi.org/10.70122/FK2/L7CNP1/SM5AIR>, Demo Dataverse, V1, UNF:6.69/RG24ZRqkmoYSG5IA== [fileUNF]

Cite Data File ▾ Learn about [Data Citation Standards](#).

Dataset Citation

Leahey, Amber, 2024, "TEST Labour Force Survey: January 2024 [Canada]", <https://doi.org/10.70122/FK2/L7CNP1>, Demo Dataverse, V1, UNF:6.69/RG24ZRqkmoYSG5IA== [fileUNF]

Cite Dataset ▾ Learn about [Data Citation Standards](#).

File Tools Metadata Versions

File Tools ▾ [Open in New Window](#)

Tell me what you want to know

Give me all the variable [labels](#)

Answer please

The variable labels in the dataframe are as follows: - REC\_NUM - SURVYEAR - SURVMNTH - LFSSTAT - PROV - CMA - AGE\_12 - AGE\_6 - SEX - MARSTAT - EDUC - MJH - EVERWORK - FPTLAST - COWMAIN - IMMIG - NAICS\_21 - NOC\_10 - NOC\_43 - YABSENT - WKSAWAY - PAYAWAY - UHRSMIN - AHRSMIN - FTPTMAIN - UTOHRS - ATOTHRS - HRSAWAY - YAWAY - PAIDOT - UNPAIDOT - XTRAHRS - WHYPT - TENURE - PREVTE - HRLYEARN - UNION - PERMTEMP - ESTSIZE - FIRMSIZE - DURUNEMP - FLOWUNEM - UNEMFTPT - WHYLEFTO - WHYLEFTN - DURJLESS - AVAILABL - LKPUBAG - LKEMPLOY - LKRELS - LKATADS - LKANSADS - LKOTHERN - PRIORACT - YNOLOOK - TLOLOOK - SCHOOLN - EFAMTYPE - AGYOWNK - FINALWT

REC_NUM	SURVYEAR	SURVMNTH	LFSSTAT	PROV	CMA	AGE_12	AGE_6	SEX	MARSTAT	EDUC	MJH	EVERWORK	FPTLAST	COWMAIN	IMM
1	2024	1	4	24	0	12		2	1	2		2			3
2	2024	1	4	47	0	4		1	6	2		3			3
3	2024	1	1	10	0	1	2	1	6	2	1			2	3
4	2024	1	1	35	4	12		1	2	4	1			2	3
5	2024	1	1	35	0	6		1	1	4	1			2	3
6	2024	1	4	59	0	10		2	2	4		2			3
7	2024	1	1	24	2	10		1	5	6	1			2	3
8	2024	1	1	59	0	3	6	1	2	4	1			2	3
9	2024	1	1	59	0	11		1	1	2	1			2	3
10	2024	1	3	48	0	5		2	1	4		1	2	2	3

Viewing rows 1 through 10 of 100278

## File Citation

Leahey, Amber, 2024, "LFS\_January\_2024.tab", *TEST Labour Force Survey: January 2024 [Canada]*, <https://doi.org/10.70122/FK2/L7CNP1/SM5AIR>, Demo Dataverse, V1, UNF:6:69/RG24IZRqkomroYSG5IA== [fileUNF]

Cite Data File ▾

Learn about [Data Citation Standards](#).

## Dataset Citation

Leahey, Amber, 2024, "TEST Labour Force Survey: January 2024 [Canada]", <https://doi.org/10.70122/FK2/L7CNP1>, Demo Dataverse, V1, UNF:6:69/RG24IZRqkomroYSG5IA== [fileUNF]

Cite Dataset ▾

Learn about [Data Citation Standards](#).

File Tools

Metadata

Versions

File Tools ▾

Open in New Window

Tell me what you want to know

Is there any missing data?

Answer please

Yes, there are missing values in the dataframe. Here is the count of missing values for each column: - AGE\_6: 87973 missing values - MJH: 45824 missing values - EVERWORK: 63454 missing values - FTPTLAST: 101213 missing values - COWMAIN: 37759 missing values - NAICS\_21: 37759 missing values - NOC\_10: 37759 missing values - NOC\_43: 37759 missing values - YABSENT: 104318 missing values - WKSAWAY: 104318 missing values - PAYAWAY: 104931 missing values - UHRMAIN: 45824 missing values - AHRMAIN: 45824 missing values - FTPTMAIN: 45824 missing values - UTOTHRS: 45824 missing values - ATOTHRS: 45824 missing values - HRSAWAY: 58206 missing values - YAWAY: 102478 missing values - PAIDOT: 58206 missing values - UNPAIDOT: 58206 missing values - XTRAHRS: 58206 missing values - WHYPT: 97294 missing values - TENURE: 45824 missing values - PREVTEN: 101213 missing values - HRLYEARN: 54251 missing values - UNION: 54251 missing values - PERMTEMP: 54251 missing values - ESTSIZE: 54251 missing values - FIRMSIZE: 54251 missing values - DURUNEMP: 105199 missing values - FLOWUNEM: 105078 missing values - UNEMFTPT: 105078 missing values - WHYLEFTO: 101213 missing values - WHYLEFTN: 101213 missing values - DURJLESS: 70686 missing values - AVAILABL: 104692 missing values - LKPUBAG: 108655 missing values - LKEMPLOY: 107461 missing values - LKRELS: 107965 missing values - LKATADS: 106212 missing values - LKANSADS: 107563 missing values - LKOTHERN: 108276 missing values - PRIORACT: 105453 missing values - YNOLOOK: 107820 missing values - TLOLOOK: 109024 missing values - SCHOOLN: 27819 missing values - AGYOWNK: 80788 missing values

REC_NUM	SURVYEAR	SURVMNTH	LFSSTAT	PROV	CMA	AGE_12	AGE_6	SEX	MARSTAT	EDUC	MJH	EVERWORK	FTPTLAST	COWMAIN	IMM
1	2024	1	4	24	0	12		2	1	2		2			3
2	2024	1	4	47	0	4		1	6	2		3			3
3	2024	1	1	10	0	1	2	1	6	2	1			2	3
4	2024	1	1	35	4	12		1	2	4	1			2	3
5	2024	1	1	35	0	6		1	1	4	1			2	3

More applications...

# Dataverse & Library AI tools

- We are seriously talking about ‘data’ as in research datasets, files, documentation, and code.
- Many data and metadata formats
  - Individual files & different types and formats
  - Metadata contents: file names, labels, categories, units, codes, questions, etc.
  - Metadata discovery: DataCite, Schema.org, croissant, and more standards
- Exposure of metadata for LLM AI to consume and preparedness for these types of services (dataset and file level)
- Need for flexibility given domain specificities and technical infrastructure interoperability (e.g. a variety of domain standards, ontologies, vocabularies, etc.)
- Need to understand current needs & future demands
  - What is the cost \$ to run something for end-users versus for internal staff procedures?

# Resources

- Borealis (<https://borealisdata.ca>)
- Harvard Demo [Test Dataset](#) 'Test LFS dataset'
- [ICPSR's TurboCurator About](#)
- [ICPSR's TurboCurator FAQ](#)
- [Odesi Metadata Best Practices](#)
- [Odesi Deposit Guide](#)
- [UVic Collection Deposit Guide](#)