

AI and The **Dataverse**[®] Project

Using AI to Benefit Research Data Community Members

Open Source Data Repository

137 installations, 38 countries

9 CoreTrustSeal repositories



Active Community

FAIR support

Metadata standards (e.g. **Croissant ML**)

Custom metadata

Citations, Embargoes, Provenance, other Features

Previewers

External tools

Interoperability/Application integrations



Harvard Dataverse Repository

Open to all researchers (not just Harvard's affiliates)
from **any discipline** (generalist repository)

All file types and disciplines are welcome

All **data** must be **de-identified**

All datasets undergo **curatorial review**

- 105K+ datasets
- 2.8+ million files
- 107+ million downloads

Dataverse Project Timeline

Dr. Gary King

Founder & Principal Investigator

Albert J. Weatherhead III University Professor

Harvard University

Institute for Quantitative Social Sciences



1997	2006	2007	2008	2015	2016	2023	2024	2025	2026
<p>Project Begins</p> <p>Focus on preserving & sharing Social Sciences data</p>	<p>The Dataverse Network 1.0</p> <p>Harvard Dataverse Repository established</p>	<p> Announcing the Dataverse Network</p>	<p>Odum Archive: First Dataverse outside Harvard</p>	<p>Data Citations</p> <p> First Open Dataverse Community Meeting</p> <p>Dataverse 4.0</p>	<p>FAIR Data</p>	<p> 100th Dataverse installation</p> <p>Dataverse 6.0</p> <p>Dataverse Re-architecture</p>	<p> Ask the Data</p> <p> Croissant</p> <p>New Dataset Types (e.g., software, computational workflows)</p>	<p>137 Dataverse Installations</p> <p>Dataverse 6.7</p> <p>Large Data Services</p> <p>Traditional Knowledge (TK) Labels</p> <p>Dataverse MCP Server</p> <p>Single-Page App Launch</p>	<p> The Next Frontier</p> <p>Locally FAIR Data</p> <p>Trusted Data Labels</p> <p>Dataverse Knowledge Graph</p>

Dataverse and AI

- Investigations & development began in 2023 at Harvard and across the Dataverse Community
- Emphasis on generative AI, LLM, RAG, human-in-the-loop
- Many resources, tools, and projects are in production or underway
- **AI Promise:** Improve data findability; metadata quality; data reusability; tool development & integration; support data exploration & curation



Photo by [Justin Ha](#) on [Unsplash](#)


Who Benefits from Dataverse AI?

- Data curators
- Data depositors
- Data seekers
- Dataverse repository users
- Data reusers
- Metadata harvesters
- Repository managers
- Repository strategists
- Software developers
- System integrators



Image by [Gerd Altmann](#) at [Pixabay](#)

Dataverse AI-Related Resources, Tools & Projects

Resource, Tool or Project	Type	Improves	Status
AI Guide	Resource	User experience	Production
Ask the Data	Tool	Data reusability	Production
Ask Dataverse	Tool	User experience	Experimental
AutoSage	Tool	Metadata quality, User experience	Experimental
Croissant 	Tool	Data findability	Production
Enhancing Dataset Metadata Project	Project (Tool)	Metadata quality, User experience	Development
GREI AI Taxonomy	Resource	Understanding of AI roles in RDM	Production
Research Data Metadata Knowledge Graph	Project (Tool)	Data findability, Data reusability	Development
Model Context Protocol (MCP) Server	Tool	Data findability, Data reusability	Production
Spam Detection & Workflow Automation	Project (Tool)	Repository data quality	Investigation
TurboCurator	Tool	Metadata quality, User experience	Production

Dataverse AI Guide

About the Guide ([link](#))

Goal: Provide Dataverse platform developers and users with access to supported AI tools and protocols

Currently Featured Tools


- Ask Dataverse
- Ask the Data
- AutoSage
- Data Chat
- TurboCurator
- Model Context Protocol (MCP) Server



Improves: User experience



Who Benefits? Dataverse repository users

 **Dataverse Guides** v6.7.1 [About Dataverse](#)

Search

User Guide

Admin Guide

AI Guide

API Guide

Installation Guide

Contributor Guide

Developer Guide

Container Guide

Style Guide

QA Guide

Search

AI Guide

Artificial Intelligence (AI) is a growing component of the Dataverse ecosystem.

Contents:

- Tools
 - Ask Dataverse
 - Ask the Data
 - AutoSage
 - DataChat
 - TurboCurator
- Protocols
 - Model Context Protocol (MCP)

Tools

Ask Dataverse

Ask Dataverse (ask.dataverse.org) is a place to ask questions about the Dataverse Project and the Dataverse software. It was created by Siava Tykhonov who [announced](#) it in December 2024 and presented it February 2025 ([video](#), [slides](#), [notes](#)).

Ask the Data

Ask the Data is an external tool that allows you ask natural language questions about the data contained in Dataverse tables (tabular data). See the [README.md](#) file at <https://github.com/IQSS/askdataverse/tree/main/askthedata> for the instructions on adding Ask the Data to your Dataverse installation.

AutoSage

AutoSage provides metadata suggestions for datasets. Learn more at <https://github.com/siacus/autosage>.

DataChat

DataChat is a multilingual open source natural language interface for Dataverse and other data platforms with an experimental Graph AI implementation for Croissant support. DataChat can literally talk back to you and explain what is inside of every single dataset, you can ask any question and it responds on the level of metadata described by Croissant standard. Learn more at <https://github.com/gdco/datachat>.

TurboCurator

TurboCurator is an external tool that generates metadata improvements for title, description, and keywords. It relies on OpenAI's ChatGPT and GPT-4 best practices. See the [TurboCurator Dataverse Administrator](#) page for more details on how it works and adding TurboCurator to your Dataverse installation.

Protocols



Model Context Protocol (MCP)

Model Context Protocol (MCP) is a standard for AI Agents to communicate with tools and services, [announced](#) in November 2024.

An MCP server for Dataverse has been deployed to mcp.dataverse.org, powered by the code at <https://github.com/gdco/mcp-dataverse>. See the code's [README](#) for information on configuring MCP clients (e.g. Cursor, Visual Studio Code, Windsurf, Zed, etc.) to use mcp.dataverse.org or your own local installation (setup instructions are also provided).

[«Previous](#) [Next»](#)

Ask the Data

 **Improves:** Data reusability, User experience
 **Who Benefits?** Data depositors


- Chatbot for tabular data files based on OpenAI
- Shiny app, 200 lines of Python code
- Created as external Dataverse tool
- Transforms NL query into SQL, executes the query on the Dataverse database, translates results into natural language
- GitHub repo: <https://github.com/IQSS/askdataverse/tree/main/askthedata>

HARVARD
DataVerse

Add Data Search About User Guide Support Cellyn Boyd

Dataset metadata of known DataVerse installations, August 2025

Version 1.0

 Gautier, Julian, 2025, "Dataset metadata of known DataVerse installations, August 2025", <https://doi.org/10.7910/DVN/RMAGSH>, Harvard DataVerse, V1, UNF:6:2N6ivs1rEVJKN6MQ0R7Ynw== [fileUNF]

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

Description This dataset contains the metadata of the datasets published in 118 DataVerse installations, information about the metadata blocks of 118 installations, and the lists of pre-defined licenses or dataset terms that depositors can apply to datasets in the 100 installations that were running versions of the DataVerse software that include the "multiple-license" feature.

The data is useful for improving understandings about how certain DataVerse features and metadata fields are used and for learning about the quality of dataset and file-level metadata within and across DataVerse installations.


How the metadata was downloaded

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

Subject Computer and Information Science

Keyword dataset metadata, dataverse, metadata blocks

Notes This dataset does not include metadata from the repositories [QDISSEI Portal](#) and [RODBUK](#), since all datasets indexed in those two repositories were published in other repositories. The purpose of QDISSEI Portal and RODBUK is to help others find data published in other repositories.

License/Data Use Agreement  CC0 1.0


Files Metadata Terms Versions


Change View Table Tree


Search this dataset...


Filter by File Type: All Access: All


1 to 10 of 142 Files


 dataset_pids_from_most_known_dataverse_installations_2025.08.25-2025.09.02.tab
Tabular Data - 75.1 MB
Published Sep 4, 2025
1 Download
7 Variables, 499760 Observations UNF:6:uL4...gpaA==

 licenses_and_terms_metadata_2025.08.25-2025.09.02.csv
Comma Separated Values - 241.8 MB
Published Sep 4, 2025
3 Downloads
MDS: 667...dea

 license_options_for_each_dataverse_installation_2025.08.29_14.58.36.tab
Tabular Data - 177.4 KB
Published Sep 4, 2025
2 Downloads
8 Variables, 333 Observations UNF:6:V6RL...W3A==

 metadatablocks_from_most_known_dataverse_installations_2025.08.29.tab
Tabular Data - 8.9 MB
Published Sep 4, 2025
1 Download
8 Variables, 16761 Observations UNF:6:5Qh...kAA==

 **Improves:** Data reusability, User experience

 **Who Benefits?** Data depositors

Tell me what you want to know

which installations have the most pids?

Answer please

The installations with the most PIDs are as follows: 1. **DANS Data Station Archaeology**: 163,458 PIDs 2. **Harvard DataVerse**: 106,285 PIDs 3. **Open Forest Data**: 79,774 PIDs 4. **JPL Open Repository**: 45,859 PIDs 5. **Borealis**: 22,960 PIDs These are the top installations based on the number of PIDs they have.

dataverse_installation_name	dataset_pid	dataset_pid_url	dataverse_collection_alias	dataverse_collection_name	dataverse_collection_type
Domus Dados	hdl:20.500.12682/rdp/136	https://hdl.handle.net/20.500.12682/rdp/136	unifesp	Domus Dados: repositório de dados de pesquisa da Universidade Federal de São Paulo	ORGANIZATIONS_INSTITUTION
Domus Dados	hdl:20.500.12682/rdp/SEEKNF	https://hdl.handle.net/20.500.12682/rdp/SEEKNF	ggesepb	Greenhouse gas emissions from sugarcane ethanol production in Bolivia	RESEARCH_PROJECTS
Domus Dados	hdl:20.500.12682/rdp/9N2XCL	https://hdl.handle.net/20.500.12682/rdp/9N2XCL	ggesepb	Greenhouse gas emissions from sugarcane ethanol production in Bolivia	RESEARCH_PROJECTS
Domus Dados	hdl:20.500.12682/rdp/2QB9YR	https://hdl.handle.net/20.500.12682/rdp/2QB9YR	ggesepb	Greenhouse gas emissions from sugarcane ethanol production in Bolivia	RESEARCH_PROJECTS

Viewing rows 1 through 4 of 492401

- Uses OpenAI
- Requires tabular data files
- NLP → SQL → NLP

Ask Dataverse

 **Improves:** User experience
 **Who Benefits?** Dataverse repository users

- Website: ask.dataverse.org
- Created by Slava Tykhonov (DANS) in December 2024
- External tool designed as "a place to ask questions about the Dataverse Project and the Dataverse software"
- Trained on Dataverse documentation
- Ask Dataverse resources: [video](#), [slides](#), [notes](#)

Available Collections

Select collection to explore

Dataverse

Collection Details

Vectors: 313,513

Vector Size: 384

Distance: Cosine

🗨️ Ask Questions

Example: How to manage permissions in Dataverse?

Enter your question:

How to manage permissions in Dataverse?



Answer

To manage permissions in Dataverse, administrators can access the Permissions page under the Edit dropdown on the Dataverse collection page. This page is divided into three sections: Permissions, Users/Groups, and Roles.

In the Permissions section, admins can decide the requirements for adding datasets and sub-Dataverse collections to their collection, as well as set permissions for users who add them.

In the Users/Groups section, admins can assign roles to specific users or groups, determining which actions they are permitted to take on their Dataverse collection. They can also reference a list of all users who have roles assigned to them and remove those roles if needed.

Additionally, admins can manage dataset permissions under this tab. Note that this setting does not retroactively apply roles to users who have previously added datasets to the collection; it only applies to users adding new datasets going forward.

In the Roles section, admins can reference a full list of roles that can be assigned to users of their Dataverse collection.

To assign roles to users and groups, admins can add, edit, or remove the roles granted to them under the Users/Groups tab. A role is a set of permissions granted to a user or group when they're using the Dataverse collection. Users and groups can hold multiple roles at the same time if needed, and roles can be removed at any time.

Sources

Title	Relevance	Source URL
Dataverse Collection Management — Dataverse.org	73.1%	http://guides.dataverse.org/en/latest/user/dataverse-management.html
Dataverse Collection Management — Dataverse.org	73.1%	https://guides.dataverse.org/en/latest/user/dataverse-management.html
Dataverse Collection Management — Dataverse.org	72.5%	https://guides.dataverse.org/en/6.2/user/dataverse-management.html

☐ Show detailed content

📄 Add Page

Enter webpage URL:



Current collection: Dataverse

☐ Crawl entire website ⓘ



 **Improves:** User experience
 **Who Benefits?** Dataverse repository users

AutoSage AI Helper

 **Improves:** Metadata quality, User experience
 **Who Benefits?** Data depositors, Data seekers

- R-shiny app
- Suggests additional dataset metadata (subjects & keywords) given a dataset DOI or its title and description
- Based on the LLAMA-2-7B and LLAMA-3.2-3B models
- Fine-tuned on the full Harvard Dataverse corpus
- GitHub repo:
<https://github.com/siacus/autosage?tab=readme-ov-file>

- **First model** (LLAMA-2-7B) trained on 76,110 datasets to June 2024
- Model shows an accuracy of 94.6%
- Details and reference to scripts:
<https://arxiv.org/abs/2411.00890>
- **Second model** (LLAMA-3.2-3B) trained on a subset of a clean and balanced training set (data balanced by subject category).
- Accuracy varies around 80%
- Paper is forthcoming

AutoSage - AI Helper

Controls

Add Subject Category

Agricultural Sciences

Add and Suggest Keywords

Select Category to Add Keyword

Agricultural Sciences

Manual Keyword

Add Keyword(s)

Enter Dataset DOI

Get Metadata From Repository

Suggest More Metadata

Reset Data

Bugs or suggestions? Visit the GitHub repo.

© 2025 S.M. Iacus (Shiny App, AI Model),
B. Treacy (AI Model)

Input Data

Dataset Title

Replication data for "Last Glacial Lake sediments from ice-free Arctic oasis reveal warm Heinrich 2 stadial summers"

Dataset Description

This dataset includes all biomarker (alkenone saturation – UK37, and alkane distribution) and geochemical (XRF – log(Br/Ti)) data presented in van der Bilt et al. 2025 – published in Nature Communications. The data derive from a lake sediment record from Svalbard – core HAP0212 (79 °N, 11 °E) – and were generated using gas chromatography (GC; alkenone and alkane data) and X-Ray Fluorescence (XRF) core scanning (log(Br/Ti)) at the geochemistry facility of the Lamont Doherty Earth Observatory (LDEO) and the EARTHlab sediment facility at the University of Bergen (UiB), respectively. All of the provided data are shown in Figs. 2, 3, 4, and 5 of the paper. Additional details are found in the readme file. (2025-05-15)

Suggest Subject Categories

Overwrite Existing Keywords?

Suggested Subject Categories

Medicine, Health and Life Sciences

Social Sciences

Earth and Environmental Sciences

Arts and Humanities

Keywords by Category

Medicine, Health and Life Sciences

Social Sciences

Health

Policy

Social Sciences

Tilburg University

Earth and Environmental Sciences

Alkenones

Alkenes

Biomarkers

Geochemistry

Lake Sediments

NE Arctic

Nordic

Northern Europe

Paleoclimatology

Paleoclimate

Svalbard

UK37

Arts and Humanities

Alkenones

Biomarkers

Climate Change

Earth Science

Geochemistry

Lake Sediments

NORWAY

Paleoclimatology

Paleoclimate

Svalbard


Improves: Metadata quality, User experience

Who Benefits? Data depositors, Data seekers


 The Dataverse Project

dataverse.org
 @IQSS

Ceilyn Boyd

Croissant 🥐

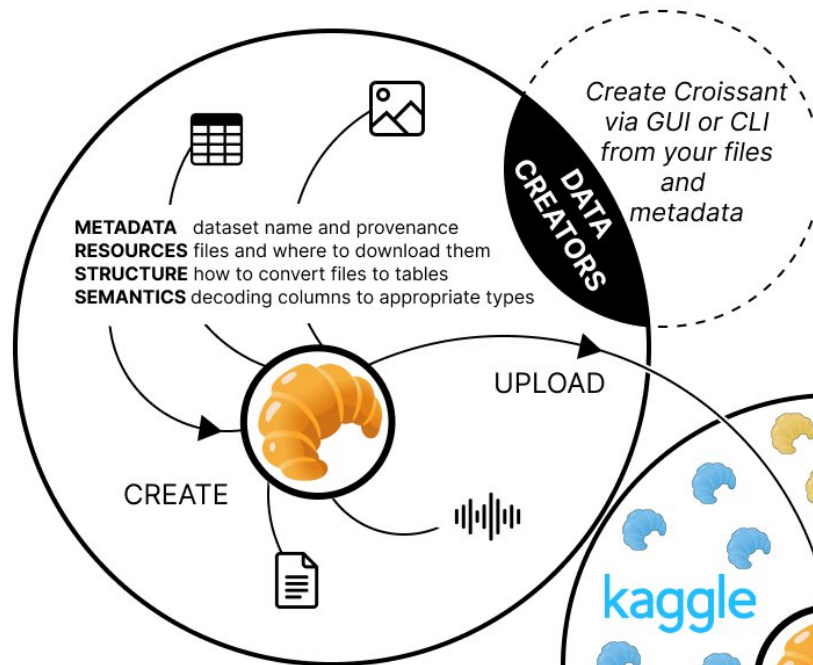


Improves: Data findability





Who Benefits? Data seekers, Metadata harvesters

- "[Croissant](#) 🥐 is a high-level format for machine learning datasets that combines **metadata**, resource file descriptions, data structure, and default ML semantics into a single file; it works with existing datasets to make them easier to find, use, and support with tools."
- Built on the [schema.org](#) standard
- Dataverse supports Croissant metadata exporting
- Findable via Google Dataset Search



The GREI Data Repository AI Taxonomy

 **Improves:** Understanding of AI roles in RDM
 **Who Benefits?** Product managers, Repository strategists

About Generalist Repository Ecosystem Initiative (GREI)

- Collaborating repositories



Goals

1. Improve researcher data sharing
2. Improve discoverability
3. Support reproducibility
4. Avoid duplication
5. Increase data reuse

About the GREI AI Taxonomy



- Classifies the distinct roles AI can play in data repository operations
- Provides a common language and reference point to foster productive discussions among stakeholders
- Provides framework to identify & address AI integration challenges in data repositories
- Chodacki, J., Hanhel, M., Iacus, S., Scherle, R., Olson, E., Pfeiffer, N., Holmes, K., & Hosseini, M. (2024). GREI Data Repository AI Taxonomy (Version 1). arXiv.
<https://doi.org/10.48550/ARXIV.2411.08054>

The GREI Data Repository AI Taxonomy

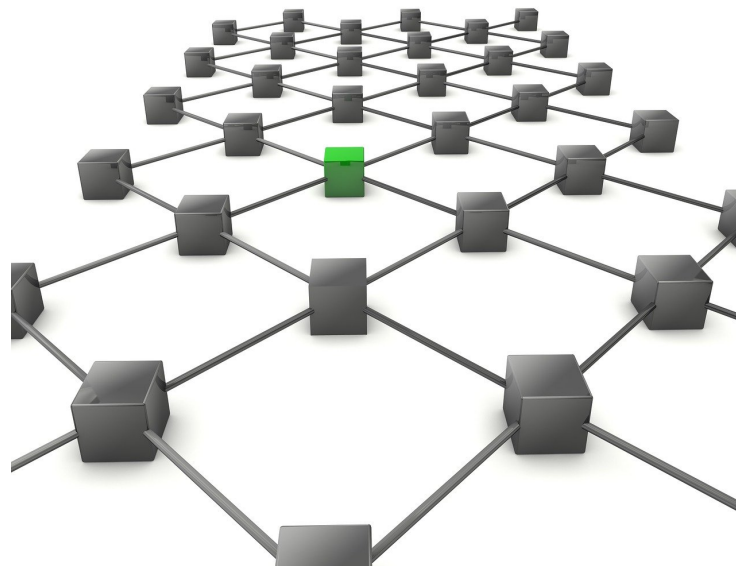


Category	Description	Example
Acquire	Gather and ingest data and metadata from different sources	AI-assisted ingest of online research data
Validate	Ensure data and metadata quality, accuracy, and integrity	AI-assisted ID of data anomalies or errors
Organize	Categorize and structure data and metadata	Tag data with keywords and descriptors
Enhance	Enrich and augment data and metadata with annotations or standard formats	Build cross-ontology knowledge graphs
Analyze	Use AI-informed analysis to uncover insights, patterns, and trends in data and metadata	AI-assisted video to text transcriptions
Share	Support data and metadata discovery, access, and distribution	AI-assisted citation generation
Support	Provide suggestions and answer questions for data and metadata users	Real-time data submission feedback

Dataverse Model Context Protocol (MCP) Server

 **Improves:** Data findability, Data reusability, Interoperability
 **Who Benefits?** Developers, System integrators

- "Open standard, open-source framework standardizing how AI systems integrate and share data with external tools, systems, and data sources"
- Dataverse MCP Server: mcp.dataverse.org
- GitHub repo: <https://github.com/gdcc/mcp-dataverse>



Peggy und Marco Lachmann-Anke from [Pixabay](https://pixabay.com/)

- TurboCurator ([link](#))
- For Dataverse 6.1+
- Implemented as a Dataverse external tool
- Improves draft and published dataset metadata
- Uses OpenAI ChatGPT to suggest metadata fields including title, description & keywords
- Leverages the [ICPSR social science thesaurus preferred terms](#)
- For details, see the use case: [TurboCurator Use Case](#)

Dataverse_1

(Solidum Producciones y Consultoria)

Demo Dataverse > Dataverse_1 >

Growth Strategies - Raw Source Data

Version 1.0



Anez, Diomar; Anez, Dimar, 2025, "Growth Strategies - Raw Source Data", <https://doi.org/10.70122/FK2/IHSWZ>
 U, Demo Dataverse, V1, UNF:6:gn+N6Y5fB8ggNaXrjd6vSg== [fileUNF]

Cite Dataset ▼

Learn about [Data Citation Standards](#).

Description ?

This dataset contains raw, unprocessed data files pertaining to the management tool group focused on '**Growth Strategies**'. The data originates from five distinct sources, each reflecting different facets of the tool's prominence and usage over time. Files preserve the original metrics and temporal granularity before any comparative normalization or harmonization.

Data Sources & File Details:

- Google Trends File (Prefix: **GT_**):
 - Metric: Relative Search Interest (RSI) Index (0-100 scale).

Access Dataset ▼

Edit Dataset ▼

Files (Upload)

Metadata

Terms

Permissions ▶

Preview URL

Thumbnails + Widgets

Configure Options ⚙️

TurboCurator by ICPSR

Deaccession Dataset

Contact

Dataset Metadata

13 Downloads

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?

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

Current Harvard/IQSS AI Projects

	Enhancing Dataset Metadata	Research Data Metadata Knowledge Graph	Spam Detection & Workflow Automation
Goal	Extract granular, variable-level metadata at scale	Develop knowledge graph of metadata and relationships for new data acquisitions	Investigate AI technologies to reduce spam and deposits missing key information
Improves	Metadata quality	Data findability & reusability	Data curation
Who Benefits?	Data depositors Data seekers Data reusers	Data seekers Data reusers	Data curators Repository managers
Status	In development	In development	Planned

Summary: Who Benefits from Dataverse AI?

Guides

Chatbots

Knowledge Graph

Metadata Tools

Metadata Standards

Protocols



Data curators

Data depositors

Data seekers

Dataverse repository
users

Data reusers

Metadata harvesters

Repository managers

Repository strategists

Software developers

System integrators



Image by [Gerd Altmann](#) at [Pixabay](#)

Obrigado!

