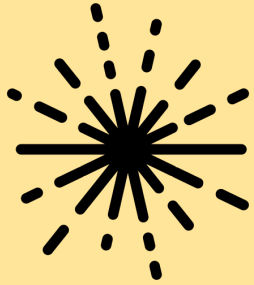


#OpenDP2020

OpenDP Use Cases

Uses cases envisioned for OpenDP

- **Archival data repositories** to offer academic researchers privacy-preserving access to sensitive data.
- **Government agencies** to safely share sensitive data with researchers, data-driven policy makers, and the broader public.
- **Companies** to share data on their users and customers with academic researchers or with institutions that bring together several such datasets.
- Collaborations between **government, industry, and academia** to provide greater access to and analysis of data that aids in **understanding and combating the spread of COVID-19**.



#OpenDP2020

OpenDP Use Cases Data Repositories

Three uses cases relevant to Data Repositories

An **OpenDP system** integrated with a data repository could be used to:

1. Enable variable **search and exploration** of sensitive datasets deposited in the repository
2. Facilitate **reproducibility** of research with sensitive datasets
3. Enable **statistical analysis** of sensitive datasets accessible through the repository

Data Repository Example: Dataverse

Dataverse Software



- Open-source software for data repositories
- 59 Dataverse repositories world-wide
- A community of users and contributors

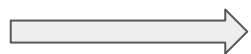
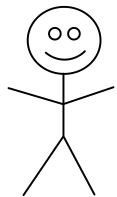
Harvard Dataverse

The screenshot shows the Harvard Dataverse website interface. At the top, it says 'HARVARD Dataverse' with navigation links for 'Add Data', 'Search', 'About', 'User Guide', 'Support', 'Sign Up', and 'Log In'. The main content area includes a call to action: 'Deposit and share your data. Get academic credit.' Below this, it displays '96,795 datasets' and '14,068,547 downloads'. A secondary call to action reads: 'Organize datasets and gather metrics in your own repository.' Below that, it shows '3,800 dataverses'. A search bar is present with the text 'Find data across research fields, preview metadata, and download files'. A featured section highlights 'COVID-19 Data Collection' as a curated collection of COVID-19 data. Below this, a 'Browse by subject' section lists various fields with their respective dataset counts: Agricultural Sciences (2,072), Arts and Humanities (1,767), Astronomy and Astrophysics (779), Business and Management (560), Chemistry (284), Computer and Information Science (1,251), Earth and Environmental Sciences (2,467), Engineering (679), Law (349), Mathematical Sciences (261), Medicine, Health and Life Sciences (3,770), Physics (980), and Social Sciences (41,240). At the bottom, there are sections for 'Datasets from journal dataverses' and 'Datasets from other dataverses', with specific examples like 'Replication data for: Patent-Based News Stocks' and 'Replication Data for: Multi-Domain Convolutional Neural Network (MD-CNN) For Radial Reconstruction of Dynamic Cardiac MRI'.

- 96K searchable datasets
- 14M file downloads
- Open to all research fields

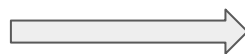
Use Case 1: Variable search and exploration of sensitive datasets

Data Depositor



Deposit sensitive data file

Dataverse[®]



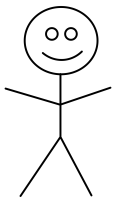
Generate DP Summary Statistics



OpenDP

1. Select privacy-loss parameter
2. Run DP query for Summary Statistics

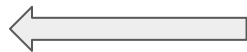
Data User



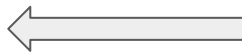
DPNA

Blue

DP summary statistics accessible



Dataverse[®]



Return DP Summary Statistics metadata file



OpenDP

Orange

Data file **not** accessible



Search and explore variable metadata and summary statistics, with privacy-protection

Use Case 1: Variable search and exploration of sensitive datasets

The screenshot shows the Harvard Dataverse interface. At the top, there's a navigation bar with 'Add Data', 'Search', 'About', 'User Guide', 'Support', 'Sign Up', and 'Log In'. Below that is a green banner for 'AJPS AMERICAN JOURNAL of POLITICAL SCIENCE'. The main content area is titled 'Replication Data for: Repression technology: Internet accessibility and state violence'. It includes a 'Use Dataset' button, a 'Cite Dataset' section, a 'Description' section with an abstract, a 'Subject' section, a 'Keyword' section, a 'Related Publication' section, and a 'Data Quality' section. At the bottom, there's a 'Files' section with a search bar and a list of files for download.

Data Explorer uses DP summary statistics generated by OpenDP

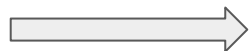
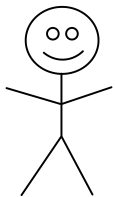
The screenshot shows the 'Data Explorer - Beta' interface. The main title is 'Replication Data for: Does Economic Inequality Drive Voters' Disagreement about Party Placement? coef.sub.all.tab'. Below the title is a search bar and a 'Results' section. A table lists variables with columns for ID, Name, and Label. To the right of the table is a 'Chart View' showing a horizontal bar chart with values for categories 6 through 10 and #N/A. Below the chart is a 'Summary Statistics' table.

ID	Name	Label
20850467	country	country
20850464	year	year
20850456	wave	wave
20850469	swldgini	swldgini
20850472	lisgini.const	lisgini.const
20850473	lisgini.linear	lisgini.linear
20850446	quintile	quintile
20850474	polity	polity
20850463	age	age

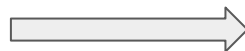
Values	Categories	N
3	6	6
6	6	17
7	7	5
8	8	84
9	9	147
10	10	436
#N/A	#N/A	25

Use Case 2: Reproducibility of research with sensitive datasets

Data Author



Dataverse[®]



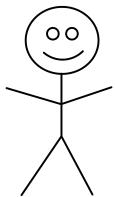
OpenDP

Deposit a sensitive
replication data file

Generate DP release
to be used to
reproduce results

1. Select privacy-loss parameter
2. Select DP query to be released

Data Reviewer
(Journal)

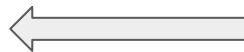


Blue

DP *replication*
release accessible



Dataverse[®]



OpenDP

Use DP release to verify
results in journal article



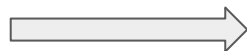
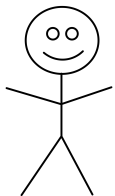
Orange

Data file **not**
accessible

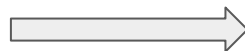
Return DP replication
release

Use Case 3: Statistical analysis of sensitive datasets

Data Depositor



Dataverse[®]



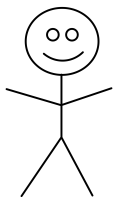
OpenDP

Deposit a sensitive data file

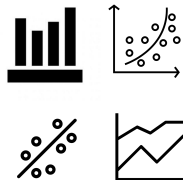
Enable DP statistics for this data file

1. Select privacy-loss parameter
2. Configure privacy-loss budget

Data User



OpenDP Interface



Dataverse[®]



OpenDP

Run any statistics using OpenDP interface, limited by privacy-loss budget



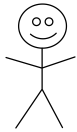
Orange 

Data file **not** accessible

Turn on OpenDP query interface for end-user

Dataverse and OpenDP integration?

Data User



Dataverse[®]

Find dataset

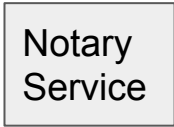


Explore data in OpenDP



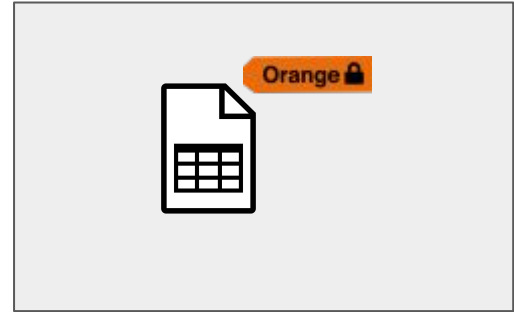
OpenDP

Use DP trusted curator model to query data

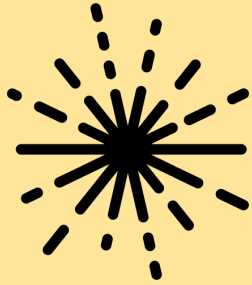


Notary Service

Trusted Remote Storage Agent Or Data Enclave



<https://cyberimpact.us/>



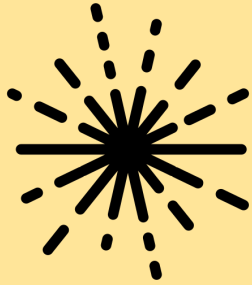
#OpenDP2020

OpenDP Use Cases Government Agencies

Use Cases relevant to Government Agencies

An increasing number of datasets useful to policy makers and researchers are being collected by government agencies. When these datasets are sensitive, OpenDP could be used to:

- Produce **rich statistical summaries** of sensitive datasets to be shared widely without risk of revealing individual-level information
- Generate **differentially private “synthetic data”** that reflects many statistical properties of the original dataset
- Allow approved researchers to use the OpenDP query interface to **run differentially private analysis of interest** on the government agency data



#OpenDP2020

Thanks

@mercecrosas