Supporting Sensitive Data in Dataverse

Dataverse Community Meeting #dataverse2020
Thursday June 18  |  9:45am -11:15am EST

AGENDA

Presentations 40m

● DataTags in Dataverse - Mercè Crosas (Harvard)
● DataTags recommendation service - Laura Huis in ’t Veld (DANS)
● OpenDP - James Honaker (Harvard)
● ImPACT: TRSA and notary service - Ilya Baldin (RENCI)

Breakout discussions & report back 30m

● Discuss use cases for sensitive data in groups

Q & A with presenters 15m
DataTags in Dataverse

Mercè Crosas, Tania Schlatter (IQSS, Harvard)
Marion Wittenberg (DANS)
Non-Sensitive vs. Sensitive data in Dataverse

**Non-Sensitive (DATAVERSE TODAY)**

- Data uploaded to Dataverse via one of the current options
- Stored locally

**Sensitive**

- Data **cannot be uploaded to Dataverse**.
- Stored in a **Trusted Remote Storage Agent**, accessed through notary service
- Metadata published in Dataverse

<table>
<thead>
<tr>
<th>Color</th>
<th>Classification</th>
<th>Accessibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Publicly open, no barriers</td>
<td>Requires Data Use Agreement (DUA); requires data enclave (moderate sensitive)</td>
</tr>
<tr>
<td>Green</td>
<td>Publicly open, but need to register to access</td>
<td>Requires DUA; stricter security requirements and audits (high sensitive)</td>
</tr>
<tr>
<td>Yellow</td>
<td>Restricted, need to be granted permissions, but non-sensitive</td>
<td>Only metadata and no link to data; data stored outside network (maximum sensitive)</td>
</tr>
<tr>
<td>Orange</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crimson</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Examples of non-sensitive vs. sensitive data sets based on Harvard Security Levels

Your institution’s Review Board determines whether the data are sensitive or non-sensitive

<table>
<thead>
<tr>
<th>Security Level 1</th>
<th>Security Level 1</th>
<th>Security Level 2</th>
<th>Security Level 3</th>
<th>Security Level 4</th>
<th>Security Level 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not need to register</td>
<td>Need to register</td>
<td>Need permission</td>
<td>Education data (FERPA)</td>
<td>Government issued identifiers</td>
<td>It would put subject’s life at risk if disclosed</td>
</tr>
<tr>
<td>Public data</td>
<td>De-identified data with low risk of re-identification</td>
<td>De-identified data with risk of re-identification</td>
<td>Datasets under contractual agreement</td>
<td>HIPAA regulated - Personal Health Information</td>
<td>Data locked in a physically secure room not connected</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Identifiable data but not considered sensitive</td>
<td>GDPR not extra sensitive level</td>
<td>GDPR extra sensitive level</td>
</tr>
</tbody>
</table>

https://security.harvard.edu/files/it-security/files/rdslexamples.pdf
Sensitive Data Support: Publishing Model

Public Repository

- Dataset and file metadata
  + Data Use Agreement (DUA)
  Set by Data Owner

Publish metadata in repository; connect metadata to sensitive data in enclave

Trusted Remote Storage Agents (TRSA) or data enclaves

- Sensitive Data Files
  Contract or agreement between Dataverse and TRSA

Trusted sensitive data depositor
Sensitive Data Support: Publishing Model

to review in breakout session

Questions:

1. How do you imagine the secure system for your Dataverse installation? Do you have own data enclaves? Do you plan to support 3rd party enclave?
   a. How might you establish the connection between the enclave and metadata?
   b. If you have a 3rd party enclave, how might you ensure that connections between the systems are secure and maintainable (technically and in terms of policy)?

2. Who do you imagine may deposit the metadata?

3. Who do you imagine may publish the metadata?

4. How to determine “sensitivity”? How do you imagine the process for your installation?
Sensitive Data Support: Data Use Model

Public Repository

Dataset and file metadata + Data Use Agreement (DUA) Set by Data Owner

Data User

Notary Service

Data access/use management; Digitally signed attestations confirming compliance with DUA & policies

Approved Secure Compute Environment

Trusted Remote Storage Agents (TRSA) or data enclaves

Sensitive Data Files
Sensitive Data Support: Differentially Private Data Release Model

Public Repository

Dataverse

Blue Green Yellow

Dataset and file metadata + Data Use Agreement (DUA) Set by Data Owner

Notary Service

Find sensitive dataset

Data User

Differentially private statistical release of the data

Secure Compute Environment to run DP statistics

OpenDP Orange Red

Trusted Remote Storage Agents (TRSA) or data enclaves

Sensitive Data Files
Sensitive Data Support: Data Use Model

to review in breakout session

Questions:

1. How do you imagine the access to the data? Is a request for access done via dataverse or via a data enclave?

2. Should there be any synchronization between the systems (metadata in Dataverse and data in the enclave)? How would it work?

3. What installations would use a tool such as OpenDP to release privacy-preserving statistics of the sensitive data?
Thursday June 18 | 9:45am - 11:15am EST

AGENDA

Presentations 40m

- DataTags in Dataverse - Mercè Crosas (Harvard)
- **DataTags recommendation service** - Laura Huis in ‘t Veld (DANS)
- OpenDP - James Honaker (Harvard)
- ImPACT: TRSA and notary service - Ilya Baldin (RENCI)

Breakout discussions & report back 30m

- Discuss use cases for sensitive data in groups

Q & A with presenters 15m
Thursday June 18  |  9:45am -11:15am EST

AGENDA

Presentations 40m

- DataTags in Dataverse - Mercè Crosas (Harvard)
- DataTags recommendation service - Laura Huis in ’t Veld (DANS)
- **OpenDP - James Honaker (Harvard)**
- ImPACT: TRSA and notary service - Ilya Baldin (RENCI)

Breakout discussions & report back 30m

- Discuss use cases for sensitive data in groups

Q & A with presenters 15m
Thursday June 18  |  9:45am -11:15am EST

AGENDA

Presentations 40m

● DataTags in Dataverse - Mercè Crosas (Harvard)
● DataTags recommendation service - Laura Huis in ‘t Veld (DANS)
● OpenDP - James Honaker (Harvard)

● ImPACT: TRSA and notary service - Ilya Baldin (RENCI)

Breakout discussions & report back 30m

● Discuss use cases for sensitive data in groups

Q & A with presenters 15m
Thursday June 18 | 9:45am -11:15am EST

AGENDA

Presentations 40m

● DataTags in Dataverse - Mercè Crosas
● DataTags recommendation service - Laura Huis in ’t Veld (DANS)
● OpenDP - James Honaker
● TRSA and notary service - Ilya Baldin

Breakout discussions 20m & report back 10m

● Discuss use cases for sensitive data in groups

Q & A with presenters 15m
Breakout discussions 20m & report back 10m

Introduce yourselves briefly, then review and discuss the questions below in your group. Add your group’s comments to this document.

How might the publishing and data use models presented apply to your installation?

1. Do you anticipate making sensitive data available in your Dataverse installation?
2. Do you anticipate having a secure system for your Dataverse installation, or relying on a remote 3rd party secure enclave Harvard’s enclave?
3. Who do you imagine may deposit the metadata?
4. Who do you imagine may publish the metadata?
5. How do you think you might determine “sensitivity” for your installation?
6. Would your installation consider using a tool such as OpenDP to release privacy-preserving statistics of the sensitive data? Why or why not?
AGENDA

Presentations 40m

- DataTags in Dataverse - Mercè Crosas (Harvard)
- DataTags recommendation service - Laura Huis in ’t Veld (DANS)
- OpenDP - James Honaker (Harvard)
- ImPACT: TRSA and notary service - Ilya Baldin (RENCI)

Breakout discussions & report back 30m

- Discuss use cases for sensitive data in groups

Q & A with presenters 15m