# 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

| | |
|---|---|
| Blue | Publicly open, no barriers |
| Green | Publicly open, but need to register to access |
| Yellow | Restricted, need to be granted permissions, but non-sensitive |

## Sensitive

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

| | |
|---|---|
| Orange | Requires Data Use Agreement (DUA); requires data enclave ***(moderate sensitive)*** |
| Red | Requires DUA; stricter security requirements and audits ***(high sensitive)*** |
| Crimson | Only metadata and no link to data; data stored outside network ***(maximum sensitive)*** |

# Examples of non-sensitive vs. sensitive data sets based on Harvard Security Levels

Blue | Green | Yellow | Orange | Red | Crimson

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

**Non-Sensitive** ⟵ ⟶ **Sensitive**

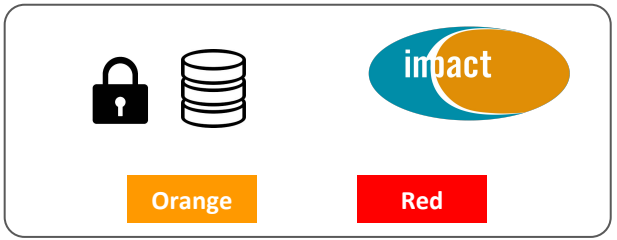
| *Security Level 1* | *Security Level 1* | *Security Level 2* | *Security Level 3* | *Security Level 4* | *Security Level 5* |
|---|---|---|---|---|---|
| Do not need to register<br><br>Public data | Need to register<br><br>De-identified data with low risk of re-identification | Need permission<br>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

https://security.harvard.edu/files/it-security/files/rdslexamples.pdf

# Sensitive Data Support: Publishing Model

**Public Repository**


**Trusted Remote Storage Agents (TRSA) or data enclaves**




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

**Sensitive Data Files**

Contract or agreement between Dataverse and TRSA

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

Trusted sensitive data depositor

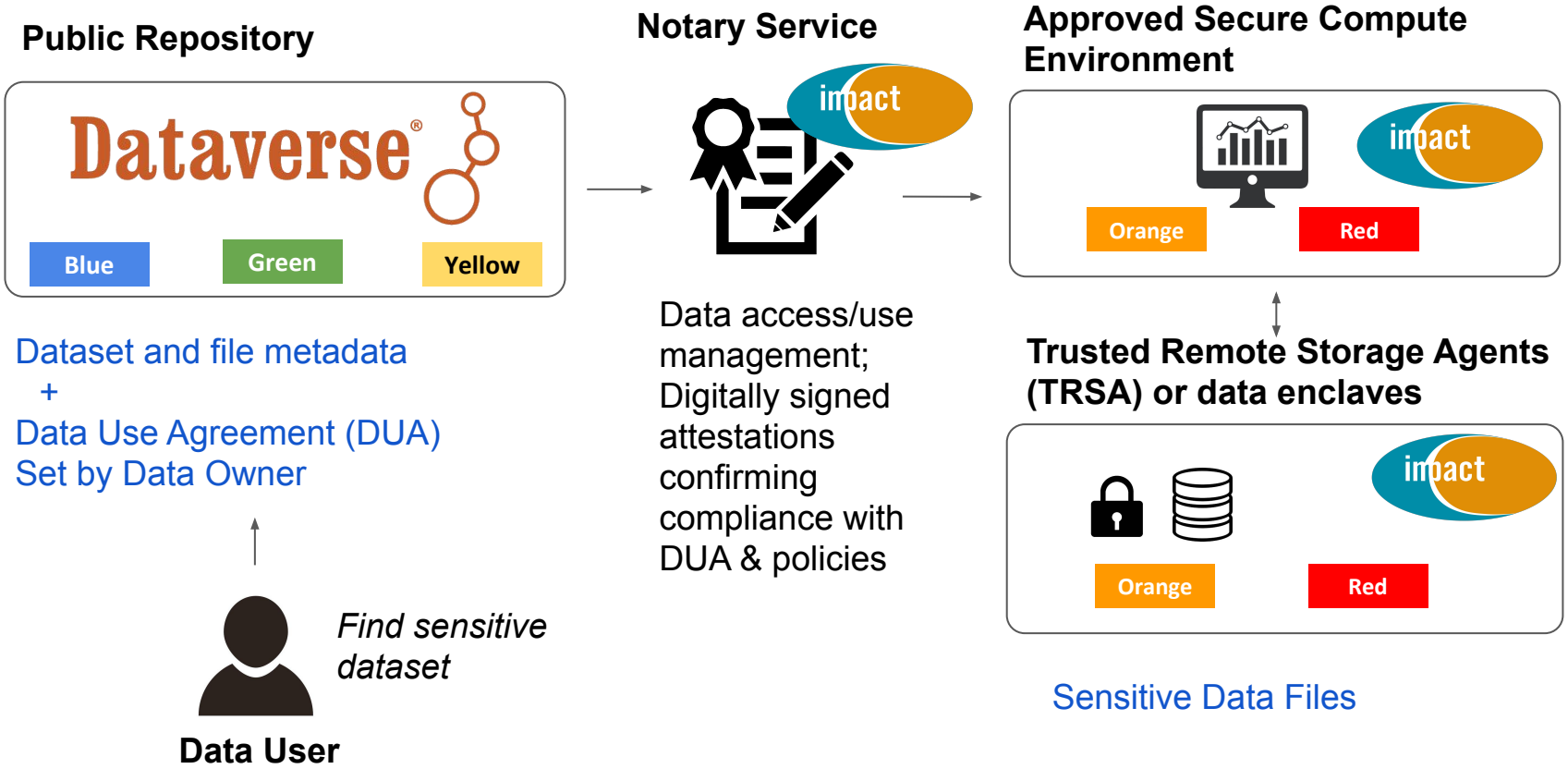
# Sensitive Data Support: Publishing Model

***to review in breakout session***

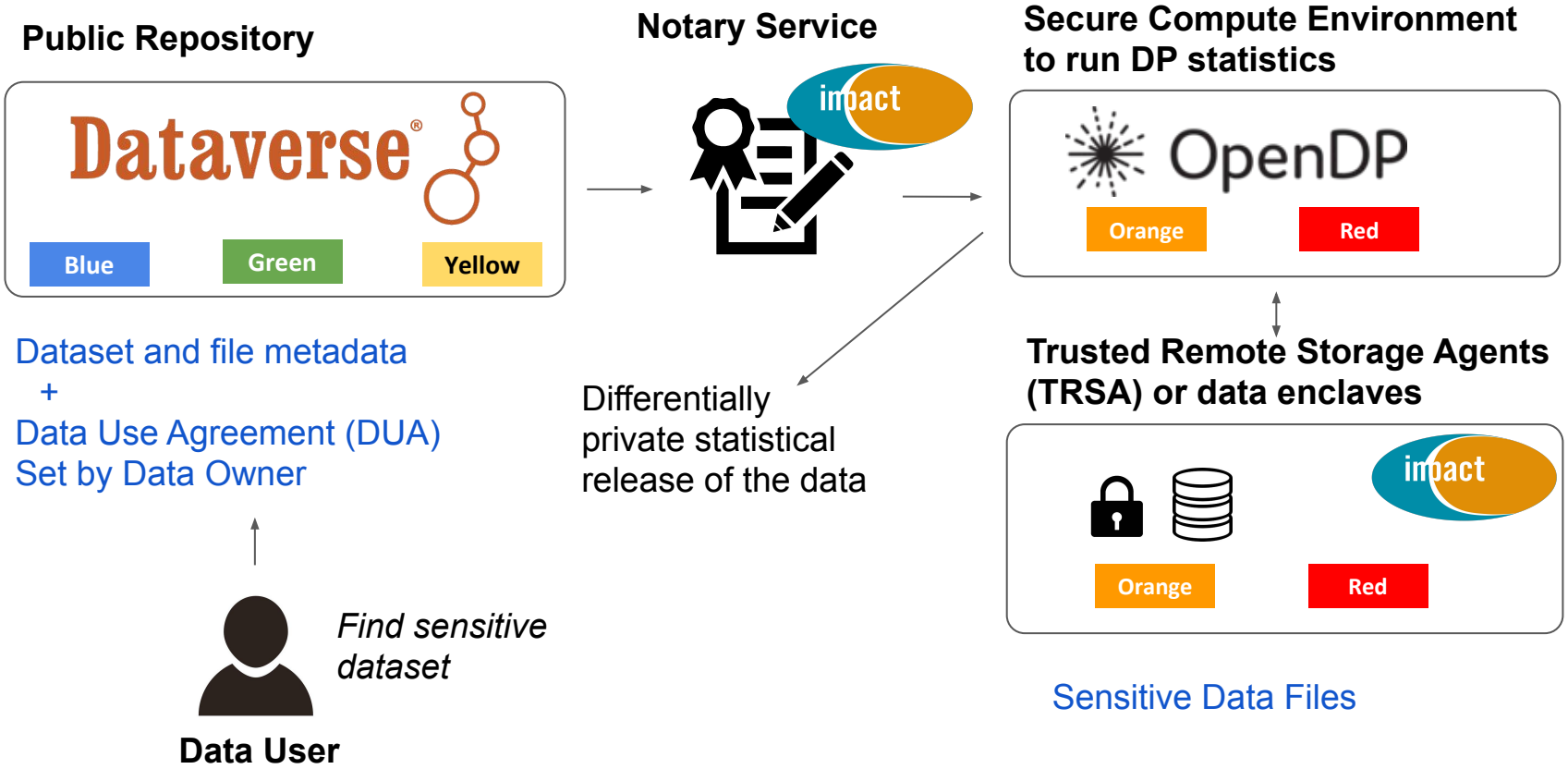
Questions:

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



# Sensitive Data Support: Differentially Private Data Release Model



# 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?

**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*