Towards an Integrated Research Data Management @Harvard

Mercè Crosas, Ph. D.
Harvard University's Research Data Officer, OVPR
Chief Data Science and Technology Officer, IQSS
@mercecrosas
Collaboration between Research, the Library, and IT is key.
Data should be trackable throughout their lifecycle.

New Harvard Agreements System tracks Data Use Agreement (DUA) for each dataset.

A reference guide with information and resources to help you manage your research data.

DATA LIFECYCLE
- Planning Data Management
- Data Acquisition and Collection
- Storage, Security, and Analysis
- Dissemination and Preservation

Dataset
- Collected by researcher
- Licensed
- Data Use Agreement

- Regular Storage or Enclaves
- Data centers or Clouds
- Notebooks Workflows Pipelines

Open Data (CCo, CCby)
Restricted, Sensitive Data (licenses, terms, Data Use Agreement)

https://researchdatamanagement.harvard.edu
Shared data should be FAIR

Harvard Dataverse facilitates making data FAIR

Dataset Landing Page must be human-readable

DataCite DOIs in data citation (Findable and Accessible)

Make Data Count (Reusable metric)

Citation and discoverable metadata using DataCite, schema.org, Dublin Core, DDI standards (Findable, Accessible and Reusable)

More metadata, including domain-specific (Reusable)

Terms with license or Data Use Agreement (Reusable)

PROV metadata (Reusable)

https://dataverse.harvard.edu
FAIR data must be machine-readable

Dublin Core meta-tags for citation metadata (Findable and Accessible)

Schema.org JSON-LD (Findable in Google Dataset Search)
Restricted data files (Authentication and Authorization needed)

Open data files (direct download)

Download data and metadata using domain-relevant standards (Reusable)

FAIR is not always Open

Landing Page with metadata, DUA, license (Accessible)
Machine-readable variable description from DDI (Interoperable and Reusable)

Summary Statistics in DDI, automatically calculated upon data upload (Interoperable and Reusable)

Metadata for each variable (or field) is needed for I and R
Data should be shared responsibly

DataTags facilitates sharing sensitive data responsibly with confidence

<table>
<thead>
<tr>
<th>Color</th>
<th>Level Description</th>
<th>Approval Needed</th>
<th>Data Protection Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Public</td>
<td>Register</td>
<td>Encrypted transmit</td>
</tr>
<tr>
<td>Green</td>
<td>Public accountable</td>
<td>Approval Needed</td>
<td>Click-thru DUA</td>
</tr>
<tr>
<td>Yellow</td>
<td>Restricted not sensitive</td>
<td>Signed DUA</td>
<td>Encrypted transmit Encrypted storage</td>
</tr>
<tr>
<td>Orange</td>
<td>Restricted sensitive</td>
<td>Signed DUA</td>
<td>Encrypted transmit Encrypted storage</td>
</tr>
<tr>
<td>Red</td>
<td>Restricted high sensitive</td>
<td>Signed DUA</td>
<td>Encrypted transmit Encrypted storage</td>
</tr>
<tr>
<td>Crimson</td>
<td>Restricted max sensitive</td>
<td>Signed DUA</td>
<td>Encrypted transmit Multi-encrypted storage</td>
</tr>
</tbody>
</table>

Sweeney, Crosas, Bar-Sinai, 2015. Sharing Sensitive Data with Confidence: The DataTags System, Technology Science
An Integrated Solution (in the near future)

Dataset

Research analysis

Reuse

DataTag + License or Agreement

Manage in repository before publish data

Publish data

DataTag + License or Agreement

Repository

The Dataverse Project