Responsible and FAIR (Findable, Accessible, Interoperable, Reusable)
Research Data Management
at Harvard University

Mercè Crosas, Ph.D.
Harvard University’s Research Data Officer, Office of Vice Provost for Research
Chief Data Science and Technology Officer, Institute for Quantitative Social Science
@mercecrosas
1. Research Data Management: becoming central to research institutions

2. Research Data Management at Harvard: collaboration is key

3. The Harvard Dataverse repository: a solution for data sharing and archiving

4. Supporting FAIR (Findable, Accessible, Interoperable, Reusable) Principles

5. The near future: a secure integrated solution
The Old Data Lifecycle

Project Planning → Data Acquisition → Data Analysis → Publication

Data Roadmap, based on K. Briney, 2015. Data Management for Researchers
The New Data Lifecycle: *Data as a Product of Research*

- Funders and Journals Data Sharing Policies
- Credit for your data

Data Roadmap, based on K. Briney, 2015. Data Management for Researchers
The New Data Lifecycle: *Data as a Product of Research*

- **Project and Data Management Planning**
- **Data Acquisition**
- **Data Analysis**
- **Publication; Data Sharing and Preservation**

**Tier 1**
- Raw data
- Data Security assigned

**Tier 2**
- Processed data
- Storage & Backup

**Tier 3**
- Published data
- Long-Term Storage

**Data Reuse**

**Reproducibility**

---

Data Roadmap, based on K. Briney, 2015. Data Management for Researchers
1. Research Data Management: becoming central to research institutions

2. Research Data Management at Harvard: collaboration is key

3. The Harvard Dataverse repository: a solution for data sharing and archiving

4. Supporting FAIR (Findable, Accessible, Interoperable, Reusable) Principles

5. The near future: a secure integrated solution
Collaboration is Key

- Policies, compliance
- Procedures, guidelines
- Best practices, outreach
- Training, support
- Repositories, tools
- Computation and storage infrastructure
A One-stop Resource

http://researchdatamanagement.harvard.edu

- Launched in January 2019
- A Harvard-wide resource
- Supports the new data lifecycle
- Links to other relevant Harvard sites
- News and Data Shares blog
- Content provided by groups across Schools, Library, IT, and Office of Vice Provost for Research
- [still growing and improving]
Working Groups to Foster Collaboration and Diverse Input

RDM Support WG
- Creates and coordinates website, use cases, DMP templates, services, best practices
- Chairs: Ceilyn Boyd, Julie Goldman
- 8 members from HL, HBS RCS, Chan Bioinformatics Core, HMS RDM, IQSS, OVPR

Electronic Lab Notebooks WG
- Coordinates ELN rollout, support and training
- Chairs: Mercè Crosas, Alan Wolf, Mason Miranda
- 6 members from OVPR, HUIT, HMS RITS, HL

Data Use Agreements Operations WG
- Develops and implements DUAs policies and procedures incoming and outgoing data
- Chairs: Jennifer Ponting, Alisa Jahns
- 16 members from OVPR, OSP, HUIT, HMS-ORA, SPH-SPA, SEAS, FAS, GSE, HMS Health Care Policy

Data Safety and Security WG
- Develops a University-wide tracking and compliance system for data security
- Chairs: Alisa Jahns, Rachel Talentino
- 9 members from OVPR, HUIT, HMS, GSE, SPH, HMS Health Care Policy
1. Research Data Management: becoming central to research institutions

2. Research Data Management at Harvard: collaboration is key

3. The Harvard Dataverse repository: a solution for data sharing and archiving

4. Supporting FAIR (Findable, Accessible, Interoperable, Reusable) Principles

5. The near future: a secure integrated solution
Open source research data repository software
A software, a community, many repositories

The Dataverse Software
http://dataverse.org
• Developed since 2006 at Harvard’s Institute for Quantitative Social Science
• 89 contributors, most external to Harvard
• > 1000 pull requests in GitHub
• 12 releases a year
• 43 installations around the world

The Harvard Dataverse Repository
http://dataverse.harvard.edu
• Open to all researchers, all disciplines
• 30,000 datasets deposited
• + 50,000 datasets harvested from other Repositories
• 250 new datasets added per month
• 7 million file downloads
A Rich Set of Features

• **Data citation:** credit as an incentive to share data
• **Metadata:** find and reuse data
  • Data Documentation Initiative (DDI)
  • DataCite (+ OpenAire)
  • Dublin Core
  • Schema.org
• **Versioning** for dataset and files
• **Tiered access to data:** guestbook, terms of use and licenses, file restrictions
• **Integration** with data exploration tools
• **Customization** and branding of your own dataverse (your collection of datasets)
• **Extensive APIs**
• **OAI-PMH for metadata** harvesting from one repository to another
1. Research Data Management: becoming central to research institutions
2. Research Data Management at Harvard: collaboration is key
3. The Harvard Dataverse repository: a solution for data sharing and archiving
4. Supporting FAIR (Findable, Accessible, Interoperable, Reusable) Principles
5. The near future: a secure integrated solution
Three Years Ago ...

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, Jusbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandro Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C.’t Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schalk, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andrea Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons

*Scientific Data* 3, Article number: 160018 (2016) | Download Citation
”The FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. ”

Wilkinson et al. 2016 “The FAIR Guiding Principles for scientific data management and stewardship” Scientific Data
The FAIR Guiding Principles

- **To be Findable:**
  - F1. (meta)data are assigned a **globally unique and persistent identifier**
  - F2. data are described with rich **metadata** (defined by R1 below)
  - F3. metadata clearly and explicitly include the identifier of the data it describes
  - F4. (meta)data are **registered or indexed** in a searchable resource

- **To be Accessible:**
  - A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
  - A2. metadata are accessible, even when the data are no longer available

- **To be Interoperable:**
  - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - I2. (meta)data use **vocabularies** that follow FAIR principles
  - I3. (meta)data include qualified references to other (meta)data

- **To be Reusable:**
  - R1. (meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards
Each dataset has a landing page with human-readable metadata and a global persistent identifier.
Each dataset has machine-readable metadata

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

Schema.org JSON-LD
(Findable in Google Dataset Search)
A dataset can contain open and/or restricted data files; Metadata is always accessible.

Restricted data files (Authentication and Authorization needed)

Open data files (direct download)

Download data and metadata using domain-relevant standards (Reusable)
Machine-readable Variable description from DDI (Interoperable and Reusable)

Summary Statistics in DDI, automatically calculated upon data upload (Interoperable and Reusable)
1. Research Data Management: becoming central to research institutions
2. Research Data Management at Harvard: collaboration is key
3. The Harvard Dataverse repository: a solution for data sharing and archiving
4. Supporting FAIR (Findable, Accessible, Interoperable, Reusable) Principles
5. The near future: a secure integrated solution
Next Challenges

• The continued growth of research data

• Complex privacy and security requirements

• Data quality: new curation services (Library, IQSS)
# DataTags Facilitate Sharing Sensitive Data Responsibly

<table>
<thead>
<tr>
<th>Color</th>
<th>Sensitive Level</th>
<th>Approval Needed</th>
<th>Security Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Public</td>
<td>Register</td>
<td></td>
</tr>
<tr>
<td>Green</td>
<td>Public</td>
<td>Register</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accountable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yellow</td>
<td>Restricted</td>
<td>Approval Needed</td>
<td>Click-thru Data Use Agreement (DUA)</td>
</tr>
<tr>
<td></td>
<td>Not Sensitive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orange</td>
<td>Restricted</td>
<td>Approval Needed</td>
<td>Signed DUA</td>
</tr>
<tr>
<td></td>
<td>Sensitive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red</td>
<td>Restricted</td>
<td>Approval Needed</td>
<td>Signed DUA Two-factor Auth</td>
</tr>
<tr>
<td></td>
<td>High Sensitive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crimson</td>
<td>Restricted</td>
<td>Approval Needed</td>
<td>Signed DUA Two-factor Auth</td>
</tr>
<tr>
<td></td>
<td>Max Sensitive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Sweeney, Crosas, Bar-Sinai, 2015. Sharing Sensitive Data with Confidence: The DataTags System, Technology Science*

Mercè Crosas, IQSS, OVPR, Harvard University @mercecrosas
Integrating DataTags with Dataverse

Data Analysis

Publication; Data Sharing and Preservation

Source Data

Incoming: DataTag + Data Use Agreement (DUA)

Outgoing: DataTag + Data Use Agreement (DUA)

Sensitive Data

Privacy-Preserving

Differential Privacy Tools

Data Analysis

Publish data

The Dataverse Project

Sensitive Data

Mercè Crosas, IQSS, OVPR, Harvard University @mercecrosas
The Harvard Data Commons: A Proposed Integrated System

Hierarchical data management:
- **Tier 1**: Raw data, Data Security assigned
- **Tier 2**: Processed data, Storage & Backup
- **Tier 3**: Published data, Long-Term Storage

Steps:
1. **Data Acquisition**
2. **Data Analysis**
3. **Publication; Data Sharing and Preservation**

Supporting tools and technologies:
- Electronic Lab Notebooks
- Big data transfer
- Policy-Based, Scalable File and Object System

Security and Access:
- DUA + Security + Access
- Provenance + code/container

Metadata and Identification:
- Persistent ID

Check-in process:
- Unpublished dataset in repository

Publish process:
- Publish dataset

Additional resources:
- Harvard Dataverse
- iRSpace
- Globus
- Starfish
Thanks!

Mercè Crosas, Ph.D.
Harvard University’s Research Data Officer, Office of Vice Provost for Research
Chief Data Science and Technology Officer, Institute for Quantitative Social Science
@mercecrosas