A Future of Research Transparency: Enabling Reproducibility in Repository Design

Mercè Crosas
Institute for Quantitative Social Science, Harvard University
mcrosas@g.harvard.edu

2019

INTRODUCTION

The proposed "Institute for the Secure Sharing of Online Data" (ISSOD) is a new initiative that aims to establish an institute to:

(a) act as a data repository for large-scale social and digital media data sets

(b) provide a replication archive for large sensitive scale social and digital media datasets

(c) establish a new "National Information Survey" that provide regular surveys to monitor trends in digital information consumption.

In this whitepaper, I will focus on addressing aims (a) and (b) above. My recommendations and review of the current landscape in these topics are mostly based on my experience leading the Dataverse project (King 2007; Crosas 2011; The Dataverse Project 2018) for more than a decade.

Addressing A: to act as a data repository for large-scale social and digital media data sets

Before planning to build a data repository, we should evaluate what *technology solutions* already exist to support such a significant effort and what *features* the data repository should provide to be up to date with current standards and best practices in data sharing, data publishing, and archiving. In this section, first, I'll evaluate options for building and hosting a data repository and then describe a set of desired repository features.

Technology options for hosting a data repository

There are three main approaches that ISSOD could take to support a data repository:

Option 1) Host the data in an existing repository that provides most of the features desired by the new institute,

Option 2) Use a current repository technology for building the new data repositories, and if needed, customize it or collaborate to develop new features, and

Option 3) build the technology from scratch to match precisely the needs of the new repository.

Let's explore these three options in the context of ISSOD's aims.

Option 1: Should ISSOD use an existing data repository instead of hosting its own? ISSOD could collaborate with one of the top existing social science repositories to host the new type of data (large-scale media data) that the new institute wants to provide for other researchers. For example, the new institute could host the data through the ICPSR repository (Inter-University Consoritum for Political and Social Research 2018), the Odum Institute repository (The Odum Institute 2018), or the Harvard Dataverse repository (Harvard Dataverse 2018), ISSOD would function as the data provider or data author—which would include acquiring or collecting the data, curating them, and deciding on the terms of use and access, but the data would be distributed and stored in one of these existing repositories. Although this would be less work and responsibility for ISSOD (with no need for technical or development resources) than building its own repository, it would also mean less control on the data, costs, and features.

Option 2: Should ISSOD use existing repository software to build the new repository? ISSOD could use an open source platform such as the Dataverse software (The Dataverse Project 2018), Globus (Sotomayor and Childers 2006), or CKAN (CKAN 2018), to build and host a repository. In this case, ISSOD would need to decide where to host the repository—dedicated servers or on the cloud—and support technical resources to install and maintain the repository in production 24/7. This option, however, would give flexibility on how and where to store the data. By choosing an open source software solution, the ISSOD could contribute to developing new features or customizations on top of the existing platform. If the open source project has an active community, the new development could be shared with other teams who have similar interests.

Option 3: Should ISSOD fully develop the software for building and hosting its repository from the ground up? ISSOD could either contract or hire developers to build its own repository and have maximum flexibility on its features. Given that existing technology exists for data sharing and data management, it would be hard to justify this option financially. For instance, developing the Dataverse software platform has taken more than 10 years, with an average of 8 developers (including UI, QA, application developers), resulting in a cost of more about 1M per year. ISSOD would not need the same amount of software development as needed by the Dataverse software platform given that it would be building only one custom repository instead of a platform that needs to support many types of repositories, but this option would still require a significant investment. ISSOD could combine options 2 and 3 to use components of existing technologies in addition to implementing its own components but would need to be aware of the implications of maintaining forked, merged components from various technologies.

Repository Features A data repository should provide features that ensure, or at least facilitate, best practices in data sharing. Guidelines for best practices in data stewardship and data sharing are described in Wilkinson et al. (2016). In the paper, the authors recommend that a data repository should ensure that its data are FAIR, that is, *findable, accessible*, *interoperable*, and *reusable*. What does this mean in terms of repository features? A repository should:

- *Provide sufficient metadata to describe a dataset so that search engines or data discovery platforms can easily find the dataset.* At least some core metadata should be in a standard format that search engines can detect and process. For example, the new Google Dataset Search expects schema.org metadata, which can be embedded in JSON-LD format in the dataset landing page of the repository. Besides, detailed variable metadata for tabular data files or metadata describing the study is needed to enable others to reuse the data. For this rich metadata, it is strongly

encouraged to find metadata standards commonly used by each specific research community. The standards might vary by research domain. Using standards maximizes interoperability and facilitates merging of multiple datasets.

- *Assign a persistent identifier to each dataset, or data file, so that the data can be referenced, accessed, and located uniquely and permanently.* The persistent identifier can be a Digital Object Identifier (DOI) registered in a global registry such as DataCite ("DataCite" 2018). For proper credit to data authors or data providers, the persistent identifier should be part of a formal data citation with attribution. Best practices on implementing data citation to build incentives for data sharing are described in the Joint Declaration of Data Citation Principles (Altman et al. 2015).

- *Allow versioning of a dataset and provide a way to find, reference, and access each version.* Datasets, unlike article publications, often get updated over time, either adding more rows, recoding variables, or simply improving the metadata, and therefore it is essential when hosting data in a repository to provide means of distinguishing and identifying each version.

- *Robust support for authentication and authorization for accessing restricted data.* Authentication should provide support for multiple types of credentials (e.g., institution sign-in, OAuth), and authorization should allow for a wide range of permissions - access data and/or metadata, change data and/or metadata, share parts publicly and restrict others -. Making data *accessible*, as defined in FAIR data, does not mean necessarily that all data will be open or public. Sufficient information, however, should be provided to find the datasets and access the data use agreement (DUA) or licenses to understand how to request access to the data if permitted.

- **Application Program Interfaces (APIs)** are essential to interoperate with other systems and be able to access the metadata in a machine-readable format. Some datasets might need to be accessed via APIs from another platform and automatically deposited to the repository again via API. APIs also facilitate integration with external tools such as data exploration or visualization tools.

A detailed comparative review of features from various data repositories can be found in the following blog: https://dataverse.org/blog/comparative-review-various-data-repositories.

Large-scale data support ISSOD is interested in providing large-scale data through the repository. This means the following considerations: 1) input/output of data might not work through HTTP deposit/download. Instead, consider tools that use other protocols, such as rsync or GridFTP. Dataverse, for example, uses rsync for large file uploads and downloads, and plans to support GridFTP in the future. Globus already uses GriDFTP. Rsync is fast but only valid for non-restricted data, while GridFTP can support sensitive, restricted data; 2) For large data files, it might be best to store the data in a cloud object store. If needed, each large dataset could be stored in a different remote trusted storage for scalability purposes. Storing the data in the cloud allows moving the data closer to computation resources instead of having to download it locally for cleaning or analysis. I will describe trusted storage below in the context of secure storage for sensitive data.

Overall concerning technology options, my recommendation is to use the Dataverse open-source software platform and adapt it to ISSOD's own purpose. It should be installed in the cloud. When considering cloud options, look at the Massachusetts Open Cloud or similar public research clouds in addition to the usual cloud vendors. One possible customization could be to integrate with new upload

tools (e.g., GridFTP from Globus Transfer) and APIs to directly deposit large data sets from other sources. Another could be building light-weight data visualizations that integrate with the Dataverse API. This would give a reasonable level of flexibility on the features since additional features could be added to the open-source software, and control over the data or agreements with data partners. This is very similar to what the Qualitative Data Repository, hosted at Syracuse University, has recently done. It has extended its repository capabilities with features specific to qualitative data. The ODUM Institute and about 35 other institutions and organizations have followed a similar approach.

Addressing B: to provide a replication archive for large sensitive scale social and digital media data sets

In this section, I will discuss four critical aspects embedded in this aim: 1) sensitive data, 2) replication, 3) archive, and 4) social and digital media data. ("large scale" is also part of the aim but reviewed already in the previous section).

Support for Sensitive Data: The repositories mentioned in the previous section, and others domain-specific repositories in biomedical areas, have worked or are working towards solutions to support sensitive data. ICPSR provides data enclaves to store, access, and work on sensitive data. Harvard Dataverse and ODUM Dataverse are working on supporting tiered access to sensitive data by integrating with the DataTags system (Sweeney, Crosas, and Bar-Sinai 2015), described below.

DataTags: The DataTags system defines six levels of access and security requirements, where each dataset (or data file) in a repository is assigned one of the six datatag levels - blue, green, yellow, orange, red, and crimson -. The repository guarantees that the access and security requirements will be applied appropriately for each level. This approach facilitates sharing sensitive datasets by standardizing the restrictions given to the data from less to maximum restriction. The blue datatag applies to open or public data, which means that a user should be able to access the data without the need to either register or agree to a DUA. In this case, the data can be transferred and stored without encryption. The green datatag applies to a dataset that has no substantial restrictions or sensitive information, but the user needs to register to access it. Green datatag would be applied to data that has been de-identified, but it carries a risk of being re-identified. By capturing information about the data user, the data provider has a record of who accessed the data in case the data are re-identified in the future. The yellow datatag applies to restricted data, in which data users must be granted permissions to access the data. A DUA applies to the data, but it can be agreed to through a simple click-through. For data associated with an orange datatag, however, the user needs a signed DUA to access the data. In this case, usually, the institution or organization representing the user and the organization representing the data provider need to have mutually agreed before signing the DUA. The red datatag is the same as orange, but additionally, access to the data requires two-factor authorization. Most HIPAA data and FERPA data would be assigned a red datatag. For this white paper, the crimson datatag, the maximum restrictive level is not relevant, and in most cases, the data at this level might need to be accessed outside any network.

ISSOD could consider applying the DataTags system or a similar system to classify the datasets they plan to share into a set of well-defined levels that guarantees the requirements corresponding to each level. This standardization facilitates establishing data use agreements with the original data provider organization and reviewing with IRB the restrictions needed for sharing or accessing the data.

Differential Privacy or other Privacy Preserving tools: An attractive extension to support sensitive data is adding tools that allow the analysis of sensitive data without having to access the raw data. The Harvard Privacy Tools project has been working on a differential privacy tool, PSI (Gaboardi et

al. 2016). The current plan is to integrate the PSI tool with the Dataverse software, but it could as well be integrated with other data platforms. In conjunction with the DataTags system, the differential privacy tool will allow conducting some preliminary analysis on a yellow, orange, or red dataset without having to go through the long and tedious (and sometimes not available) process of DUA approval to be granted access to the raw data. It will also allow constructing open differentially private metadata set for a sensitive dataset, including differentially-private summary statistics.

Trusted Remote Storage Agents:

For some data sources, the organization responsible for the data might not agree to move the data to the repository's storage. A solution to this can be offered by the concept of trusted remote storage agents. To support this, a DUA would need to be signed between the organization and the repository agreeing that the data are kept in a remote storage owned and maintained by the organization, while the repository holds the metadata describing the dataset and the persistent unique URL to access the data. Both the repository and the owner of the remote storage would manage together granting access to the data (via passing a token when credentials are proven or other means).

Governance, Access Review, DUAs:

Besides the necessary technology and features to support sharing sensitive social and digital media data described above, in most cases, the challenge will be deciding on the DUA between the data source and the research institution that wants to use the data. A way to facilitate this step is to standardize, to the extent possible, the DUA and start with a template at the beginning of the negotiations.

Learning from other research communities with experience defining agreements for sensitive data could be valuable. In particular, from the biomedical community, which has established well-defined practices for sharing sensitive data for several years with projects such as TOPMed and DBGap for genomic data (https://www.nhlbiwgs.org/). In these projects, the data can be accessed for only very specific and approved research purposes.

For ISSOD, also, there are many open questions when planning to use non-public social media data for research: do social media users need to provide informed consent to allow using their data for research? Should the dataset be modified if a social media users delete a published post, following the recent European General Data Protection Regulation (GDPR)? Is it possible to guarantee that the data will be removed and synched correctly?

Replication Replication of an empirical study (or reproducibility, as it is also often referred) can be supported at various levels of complexity. At a minimum, any data and code used in the study must be made available to enable computational reproducibility of the original results. Unfortunately, in most cases, the data and code shared by the data authors do not have sufficient information to be able to reproduce the published work. This problem is often experienced by the Odum Institute at the University of North Carolina, which serves as a third-party peer reviewer to verify that the results in a submitted manuscript can be reproduced by using the data and code provided by the authors (see more on this initiative at the Odum Institute site: http://cure.web.unc.edu/odum/). On average, the Odum team needs to contact back the authors three times to gather more information before the code runs appropriately and reproduces the results. It is hard, therefore, to reproduce the results automatically if the data and code are not well formatted, reviewed, and documented.

Computational reproducibility could be improved by either using dedicated resources to curate the data (and code) once is in the repository or adding tools and policies that increase the chances that all the necessary information will be provided. The IQSS and Odum teams are working towards integrating the Dataverse platform with replication tools, such as Code Ocean or Jupyter Notebooks,

which will allow running the code on the data in an online platform, avoiding setting up a local environment. This collaboration aims to empower not only reviewers but the entire research community, to more easily be able to reuse a dataset and reproduce previous work. Once the results are verified, a 'reproduced' certification or badge should be assigned to the dataset.

Sensitive data present an additional challenge for reproducibility. If reviewers are assigned to reproduce the results in a manuscript, they will need to have the necessary permissions to access the data and code for verifying the results. One option could be to grant access to reviewers only for running the code but not allowing them to use the data for any other purposes. The DUA should include the appropriate language to enable access for peer-review. Another option, when possible, could be to integrate the replication tools with differential privacy tools, that is, provide a differentially private version of the code.

Archive Based on archival best practices, it is recommended to have three copies of the data in separate locations. Files and metadata should adhere to accepted standards or formats widely used by the research community. If data files are stored in proprietary formats or databases, the archive should provide a conversion to preservation formats to ensure they can still be used years from now. An archive should also have a succession plan, that is, a policy that defines where the data will be ported and continued to be made accessible if the archive ceases to exist in the future.

Social and Digital Media Data It would be useful to integrate the repository with network data visualizations and geospatial visualizations to support social media data. This integration would be doable through the repository API, given that the data and metadata can be accessed in a format compliant with the visualization tools. As an example of such integrations, the Harvard Dataverse integrates with the WorldMap platform (`http://worldmap.harvard.edu/`) to visualize geospatial datasets. For qualitative data (text, images, videos), the Qualitative Data Repository (`https://qdr.syr.edu/`) hosted at Syracuse University is implementing an open annotation tool to annotate individual data posts and export the annotations for analysis.

Conclusions

The ISSOD is starting a herculean task aiming to act as a data repository and replication archive for large-scale, sensitive social and digital media data. My main advice is to avoid developing the repository and archive from the ground up. Instead, use existing technologies and establish collaborations that facilitate extending the features of the repository. Then, focus your efforts and resources to acquire, clean, and curate or annotate the data to make them useful research products, as well as to build a data governance body that can help establish policies and make decisions on DUAs, set appropriate

access and security levels for the datasets, and review granting access to researchers.

References

Altman, Micah, Christine Borgman, Mercè Crosas, and Maryann Matone. 2015. "An Introduction to the Joint Principles for Data Citation." *Bulletin of the American Society for Information Science and Technology* 41 (3): 43–45. doi:10.1002/bult.2015.1720410313.

CKAN. 2018. "Open Source Data Web Portal." Accessed June 12, 2019. https://ckan.org/.

Crosas, Mercè. 2011. "The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data." *D-Lib Magazine* 17 (1/2). doi:10.1045/january2011-crosas.

"DataCite." 2018. Accessed June 12, 2019. https://datacite.org/.

Gaboardi, Marco, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. 2016. "PSI (Ψ): A Private Data Sharing Interface." Accessed May 2, 2019. arXiv: 1609.04340 [cs.SR]. http://arxiv.org/abs/1609.04340.

Harvard Dataverse. 2018. "A Data Repository for Sharing, Citing, and Archiving." Accessed June 12, 2019. https://dataverse.harvard.edu/.

Inter-University Consoritum for Political and Social Research. 2018. "Curation Services & Tools." Accessed June 12, 2019. https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/tools.html.

King, Gary. 2007. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." *Sociological Methods & Research* 36 (2): 173–199. doi:10.1177/0049124107306660.

Sotomayor, Borja, and Lisa Childers. 2006. *Globus Toolkit 4: Programming Java Services.* OCLC: ocm63805614. San Francisco: Morgan Kaufmann.

Sweeney, Latanya, Mercè Crosas, and Michael Bar-Sinai. 2015. "Sharing Sensitive Data with Confidence: The Datatags System." *Technology Science.* Accessed June 12, 2019. https://techscience.org/a/2015101601.

The Dataverse Project. 2018. "Open Source Research Data Repository Software." Accessed June 12, 2019. https://dataverse.org/home.

The Odum Institute. 2018. "Management & Curation." Accessed June 12, 2019. https://odum.unc.edu/archive/managementcuration/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. doi:10.1038/sdata.2016.18.