Reproducibility with Microsoft Research Open Data

Vani Mandava
One Microsoft Way
Redmond, Washington 98052
vanim@microsoft.com

Abstract

Access to repositories with open data sources is critical for reproducibility of research. Microsoft Research Open Data is a unique initiative that combines features of a traditional data repository with easy access to compute resources. The main aim is to increase reproducibilty of research outcomes by making datasets associated with research papers published by Microsoft researchers available broadly. Such datasets are hosted along with relevant metadata to makes it easier to discover related assets that aid reproducibilty. The repository is hosted in the cloud and has seamless integration to Azure compute enabling users to run experiments on the data using convenient cloud compute resources, thus alleviating the signifiant costs and bandwidth constraints incurred in moving the data. The reproducibilty aspect within Microsoft Research Open data is framed using three pivots - Investment, Incentive and Infrastructure.

Background

During the time between 2013 to 2017 that Microsoft Research ran cloud outreach programs in academia, we recieved hundreds of proposals and requests to support and engage in groundbreaking research across various domains. Every single such proposal had an element of discovery based on data-intensive science [1]; a remarkable evidence of Jim Gray's fourth paradigm.[1] Reproducibility was a huge concern when evaluating these proposals. Reproducibility relies on the sustainance, provenance, ease of reuse, interoperability, and accessibility of the data that the research is based on. Concurrently, there was also a desire and need to make available the datasets referenced in numerous publications from within Microsoft Research itself. Microsoft Research Open Data [2] was launched as an initiative to serve as a unified data repository for such datasets. The repository was built on Microsoft's Azure cloud infrastructure. As a natural extension of being hosted in the cloud, the idea of having not just a static data repository, but a living compute environment that enabled reproducibility emerged as a central goal for the project.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]Specifically, the first three paradigms were about research based on experimental, theoretical and computational science. The fourth paradigm was about research based on an "exaflood" of observational data that would need a new generation of scientific computing tools to manage, visualize and analyze the data flood.

Components of the Repository

The repository has the following key components

- A simple user interface that renders across devices and allows for basic search and filtering capabilities based on domain, file type, license and keyword based search as shown in Figure 1.

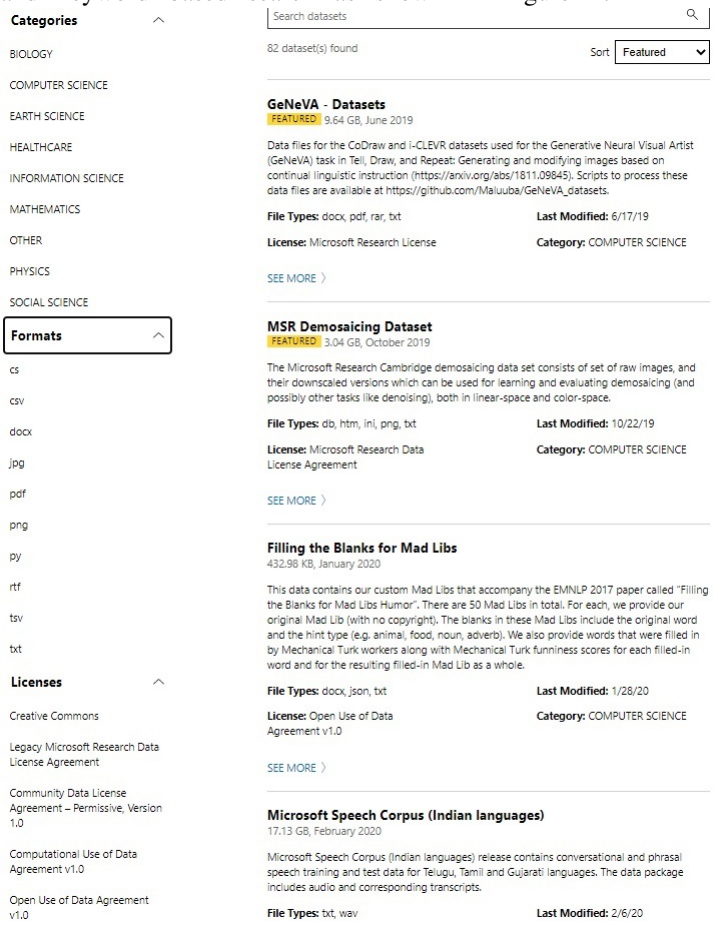

Figure 1: Microsoft Open Data Search and Filter features

- Support for both Azure B2B and B2C authentication as shown in Figure 2. This enables a separate experience for someone who owns a dataset in the repository and is signing on with an organizational ID. As a result, researchers can administer their own dataset, instead of relying on an administrator, thus keeping the repository efficient to run.

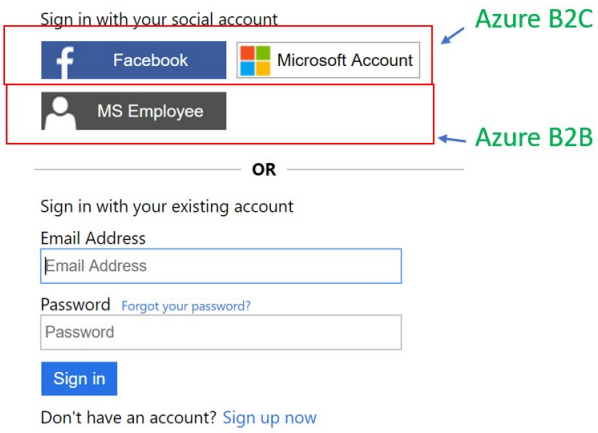

Figure 2: Authentication mechanisms

- A built in preview feature for visualizing tabular data as shown in Figure 3. Certain datatypes like csv and tsv can be visualized without needing to authenticate and download the data. This allows the user a quick view into the data at the time of browsing.

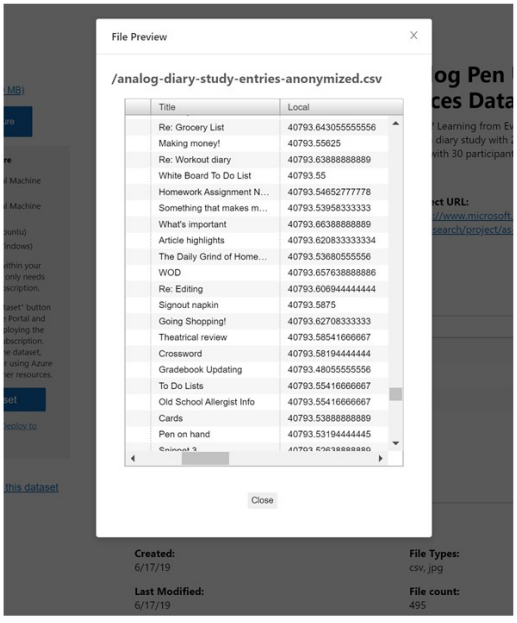

Figure 3: Preview for tabular data

Reproducibility of Research

The Microsoft Research Open Data repository is framed as a reproducibilty asset using three pivots:

Investment

Reproducibility is rarely an implicit part of the open research ecosystem. It needs planning and investment, otherwise it risks becoming an afterthought. Additionally, there is a timing aspect to reproducibility investment i.e. it is best to consider reproducibility when the research is current and active. It is harder from both effort and cost perspectives to create reproducible assets from archived research assets. There are already dozens of data repositories that researchers use. Moreover, in certain communities, there are efforts to register [3] the myriad repositories available. As a result of these prior investments, there is a risk of duplication of investment. The specific investment in Microsoft Research Open Data is a commitment to evolve how datasets curated by researchers at Microsoft are available i.e., directly via cloud compute resources. This is an important distinction that keeps data in the cloud to overcome networking and bandwidth constraints that limit data movement. Technically, this is achieved by providing a workflow where a user can a execute a cloud-to-cloud instantiation of the dataset into a data science virtual machine that comes loaded with powerful tools and libraries [4] for working with the data. In addition to making Microsoft Research datasets available via the repository, we have also made Microsoft Research Open Data available as repostitory-as-a-service. The repository is available for others to instatiate, such that they can host their own datasets, code and other reproducible assets in the cloud.

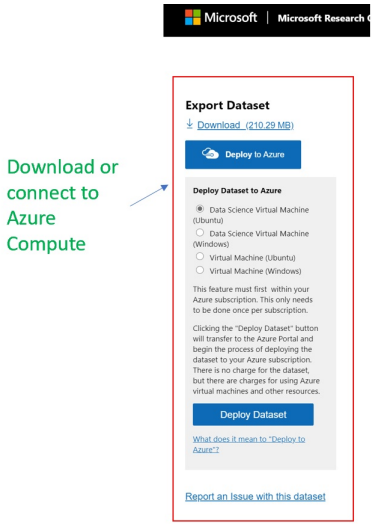

Figure 4: Computing on data in the cloud

Incentive

Reproducibility is often overlooked in the publish or perish ecosystem. There need to be incentives in place for researchers to consider the reproducibility aspect of their research. Researchers typically have many assets related to the dataset, that go beyond the dataset itself and aid in a larger understanding of the research. The meta data tag of a dataset page on the repository points back to the researcher's project or to some URL chosen by the researcher. The meta data field allows a user to browse these additonal assets related to the dataset. These assets may be hosted on a separate project page, github etc and expands the outreach and reproduciblity of the research associated with the dataset. This flexibility is an incentive to make yet another channel available to researchers to communicate their research with the community.

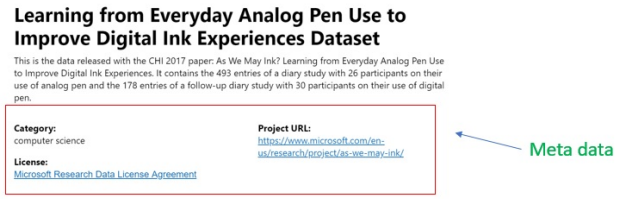

Figure 5: Metadata about the dataset

Infrastructure

Cloud based infrastructure from Microsoft Azure is the foundation of the simple platform and ease of use that Microsoft Research Open Data provides. In addition to the ability to seamlessly use cloud compute resources against datasets in the repository, the platform also leverages powerful Azure components such as CosmosDB for user metadata management, and Key Vault for encryption. It is built using the open source angular web framework that enables an elegant yet powerful user interface that allows cross device access. In addition, Azure App Insights provides usage analytics.

The user experience for the portal has distinct experiences for **users** who are browsing and using datasets, **contributors** who nominate and update datasets and metadata, and **administrators** who can execute meta level tasks such as adding new license options or updating access controls.

Data Use Agreements

The practice of data sharing is closely aligned with the ethics and legal aspects of sharing the data. This implies constantly evolving our standards for data sharing to be consistent with what was agreed to when collecting the data, and to ensure that users have a clear understanding of what they agree to when using the data.

Microsoft Research Open Data supports several data use agreements and allows data owners to provide appropriate licenses that are relevant to their dataset. Primarily, the repository aligns with two modern data use agreements [5], specifically the Open Use Data agreement (O-UDA) and the Computational User Data Use agreement (C-UDA).

The O-UDA allows distribution of the data for unrestricted uses with no privacy or confidentiality concern. It is not applicable to data protected under HIPAA or GDPR. The C-UDA, on the other hand is used when the dataset contains third party copyrighted materials. The usage of datasets under C-UDA is restricted to research use.

Acknowledgements

Microsoft Research Open Data is an outcome of the Microsoft Research Data Science Outreach program and was made possible by a collaboration between several teams at Microsoft, Microsoft researchers, industry partners, and academic advisors.

This paper was written for archival purpose from an invited talk at the AAAI 20 Reproducibility workshop on February 7th 2020, AAAI 20, NYC.

References

[1] Tony Hey, Stewart Tansley, and Kristin Tolle. The fourth paradigm: Data-intensive scientific discovery. https://www.kdnuggets.com/2009/12/pub-4th-paradigm-jim-grey.html.

[2] Microsoft research open data website. https://msropendata.com/.

[3] Heinz Pampel, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf. Making research data repositories visible: The re3data.org registry. *PloS one*, 8:e78080, 11 2013.

[4] What tools are included on the azure data science virtual machine? https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/tools-included.

[5] Microsoft research open data project: Evolving our standards for data access and reproducible research. https://www.microsoft.com/en-us/research/blog/microsoft-research-open-data-project-evolving-our-standards-for-data-access-and-reproducible-research/.