A Review of Data Repositories for the Long Tail of Computer Science

Pachev Joseph, Victor Potapenko, Naphtali Rishe, Oliver Ullrich
HPDRC, School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA

ABSTRACT

Computer scientists often struggle finding or generating data to formulate or test hypotheses, validate models, or test algorithms. This leads to greater time and effort allocated to searching for or producing data, rather than performing scientific research itself. This data barrier is especially cumbersome in the long tail of computer science – smaller laboratories typically without access to larger institutions' data sources.

This survey examines nineteen existing data repositories (see Figure 1) based on their feature sets. Out of these reviewed repositories, only six have feature sets that are different from standard digital libraries. No single data repository provides a combination of features and tools geared towards the long tail of computer science, and none offers features that would allow for purchase or sale of data. Studies (see [1] and [2]) show that providers of data in the scientific domain offer it for free approximately 80% of the time.

BABS, CCITK, CIESIN, CKAN, Clarin, Dataverse, Figshare, GEON, GitHub, IDD, OLAC, Pachyderm, PREDICT, SF, SNAP, UA-CR, UCI, Unidata, Zenodo.

Figure 1: Data repositories surveyed

BACKGROUND

A data repository is a shared data storage resource which holds many types of data to be used for analytical purposes (see [3]), providing users at least with means to upload, manage, search, and download data sets. Some of these platforms provide more advanced functions, e.g. tagging, querying, versioning, and code integration.

Studies (see [2] and [3]) focused on data providers across a variety of fields demonstrate a market in a stable and highly innovative phase, which is still dominated by vertical relations with lack of intermediaries indicating limited market efficiency. Similar conditions are apparent in the long tail of computer science, where most discoveries are made in a large number of smaller, silo-like laboratories by scientists, who have no particular incentive or specialized platform to share their data with the rest of the scientific community. Unlike most data domains that are distributed over a variety of different business models, scientific data is distributed only through two out of ten major distribution channels: marketplaces and search engines (see [2]). This indicates that scientific data is rarely sold as a standalone product.

The current trends point towards domain-focused, self-generated, specialized data (see [2]). These trends are well aligned with needs of the computer science community. There is a need to create an ecosystem that allows its participants to organize into communities, create, curate, and interlink their own sub-repositories, integrate data with code, trace data and code evolution via dataset versioning, publish and subscribe to near real-time data feeds, access the data via industry standard APIs, effectively manage licensing, and trade the data based on its value (see [2] and [3]).

EXISTING REPOSITORIES

Currently existing data repositories address widely varied data requirements of commercial and non-profit institutions, as well as the individual researchers in the field of computer science. While some provide data directly relating to the field of computer science, such as machine learning data sets, data encryption, or operating systems, others provide multidisciplinary datasets, but are geared and tooled specifically towards users that come from a computer science background and require advanced features. Figure 2 shows the six out of nineteen data repositories reviewed that provide at least basic functionality supporting the requirements of computer scientists (see Figure 3).

Repository	Data Communities	Data Marketplace	Code Integration	Versioning Control	Live Datasets
Figshare	✗	✗	✓	✗	✗
Zenodo	✓	✗	✓	✗	✗
Unidata	✗	✗	✓	✗	✓
GitHub	✗	✗	✓	✓	✗
CKAN	✗	✗	✓	✗	✗
CITK	✗	✗	✓	✗	✗

Figure 2: Comparison of selected data repositories

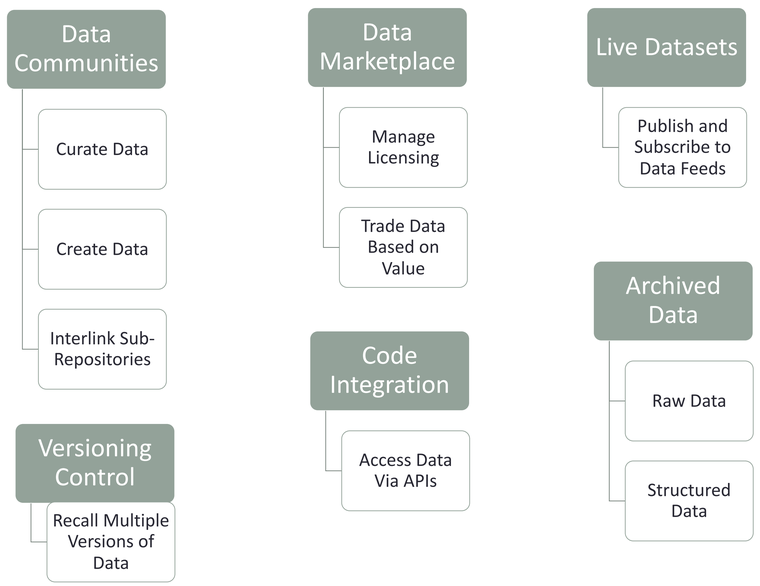

Figure 3: Requirements for data repositories as described by computer scientists surveyed in [1] and [2]

A CLOSER LOOK

The following repositories contain features especially geared towards supporting computer science research (see[1] and [2]):

- **Figshare -** A general-purpose cross-disciplinary data repository that houses more than 500,000 datasets. Figshare supports access via a REST API. Figshare is missing the support for live updates, data versioning, data market place, and community based data curation.
- **Zenodo -** Although similar to Figshare, this service is organized around the concept of communities. The service is implemented using Invenio, a free open-source digital library. Much like Figshare, Zenodo is missing the means to trade data, live dataset updates, data versioning mechanisms, and data market.
- **Unidata** - A community of 260 universities sharing tools to disseminate near real-time earth observation data online. What sets Unidata apart from the other repositories surveyed is “live datasets”: the ability to update datasets in real-time.
- **CKAN** and **CITK** - Both data repository platforms that focus on research data and software code integration, positioning themselves as toolkits for researchers. Both strive to provide data and code integration, however, they are missing dataset versioning functionality, which is key to building a useful research collaboration and data trading platform. There is no support for communities, workspaces, or commercial exchange markets.
- **GitHub -** On the opposite extreme of the dataset/code spectrum is the widely-used GitHub code repository. The repository is built using Git open-source software, created to share, track, manage and execute software projects. GitHub does not have the facilities to store large amounts of data. Git technology is built for line-by-line code versioning over a large number of individual files, and is not applicable for data repository purposes.

CONCLUSION

None of the platforms surveyed is successful at integrating all the features needed by scientists at the long tail of computer science. They are especially failing to create a marketplace where computer scientists are willing to share their own data, evaluate and provide feedback on the data submitted by others, and pay a fair price for licensing rights to the peer-reviewed data. Such a platform would enable market participants to add value to original datasets by creating algorithms that derive versions of originals, which can be used at higher levels of data analysis, while appropriately crediting the original dataset.

REFERENCES

[1] F. Stahl, F. Schomm, G. Vossen, and L. Vomfell, “A classification framework for data marketplaces,” Vietnam Journal of Computer Science, vol. 3, no. 3, pp. 137–143, 2016.
[2] F. Schomm, F. Stahl, and G. Vossen, “Marketplaces for data,” ACM SIGMOD Record, vol. 42, no. 1, p. 15, 2013.
[3] M. Assante, L. Candela, D. Castelli, A. Tani, “Are Scientific Data Repositories Coping with Research Data Publishing?,” Data Science Journal, vol. 15, 2016.