Examining Repositories for Simulation Data

Oliver Ullrich[1]*, Victor Potapenko[2], Naphtali Rishe[2]

[1]Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany, **oliver.ullrich@iais.fraunhofer.de*

[2]School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA

SNE 30(2), 2020, 61 - 66, DOI: 10.11128/sne.30.tn.10513
Received: February 26, 2020; Revised May 21, 2020;
Accepted: May 25, 2020
SNE - Simulation Notes Europe, ARGESIM Publisher Vienna,
ISSN Print 2305-9974, Online 2306-0271, www.sne-journal.org

Abstract. Researchers and practicioners often struggle finding or generating adequate data to design, calibrate, or validate simulation models. 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 review examines nineteen existing data repositories based on their feature sets. Out of these reviewed systems, only six have advanced feature sets that are significantly different from standard digital libraries. No single data repository provides a combination of features and tools geared towards simulation projects conducted at smaller laboratories, and none offers features that would allow for purchase or sale of data.

Introduction

Research projects to advance modelling and simulation methods often depend on the availability of suitable and reliable data to design, calibrate, and validate models. Significant time and effort is spent on finding or collecting, assessing, licensing, and pre-processing such data sets even before the actual modelling and simulation project can begin. This data barrier is especially cumbersome in the long tail of computer science – smaller laboratories typically without access to larger institutions' data sources.

This review examines nineteen existing data repositories (see Figure 1) based on their feature sets and utility to both data owners and data users, and with specific consideration of the long tail of computer science.

These results could also be seen as a call to action: There is definitely room for a data repository and arbiter platform, incorporating functionality aimed at supporting computer scientists in the long tail of the field, promising honorary and even financial motivation for data owners to curate and share their research data.

The paper continues with sharing some background on data repositories (see Section 1), followed by a discussion of desirable functionality, and specificially on features desired by both data owners and/or data users (see Section 2). The paper then goes on to discuss the 19 examined platforms and takes a closer look at six of the most promising of them (see Section 3). It closes with a short summary of the lessons learned an an outlook on further research (see Section 4).

1 Background

A *data repository* is a shared data storage resource that holds multiple types of data to be used for analytical or modelling purposes (see [1]), providing users at least with means to upload, manage, search, and download data sets. Some of these platforms provide more advanced functions that might include tagging, querying, versioning, and code integration (see Figure 1).

Studies (see [16] and [1]) focused on data providers across a variety of fields demonstrate a market in a stable and highly innovative phase that is still being dominated by a high vertical integration with lack of intermediaries indicating limited market efficiency. Similar conditions are apparent in the long tail of computer science, where scientists who have no particular incentive or specialized platform to share their data with the rest of the scientific community make most discoveries in a large number of smaller, silo-like laboratories.

The current trends point towards domain-focused, self-generated, specialized data (see [16]). 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 [16] and [1]). At the time of writing, it is estimated (see [17] and [1]) that providers of data in the scientific domain offer it cost free approximately 80% of the time, so that methods and tools to widely share data for honorary purposes, for example acknowledgements or co-authorships, are needed as well.

2 Potential Services

Consider this scenario: A team has developed an idea, a software tool or a new simulation technique or model to be tested. But how can the team get data to calibrate and validate these models? Where can large datasets required to evaluate the software tool be found? Once aquired, can the whole data set be included in a publication?

To solve these issues a platform or service would be needed aimed specifically at the needs of computer science researchers, motivating *data owners* to further disseminate already existing, valuable data. Such a system would enable them to monetize data sets and/or to get proper honorary acknowledgment to the data producers and their sponsoring agencies and programs, as well as to benefit from references to research papers and patents resulting from data access.

On the *data users'* side, all computer scientists are potential users of such a repository system. Especially smaller research groups without access to large data producing facilities would gain access to curated, diverse, vast amounts of data.

In the next few paragraphs some advanced functionality beyond simple data up- and downloads is envisioned that would facilitate data sharing and availability.

2.1 Services to Data Owners

An ideal system would provide data owners with a web based, encryption-enabled interface allowing researchers to deposit domain-specific datasets, to create citable Digital Object Identifiers (DOI), to define sample datasets, and to store relevant meta-data on the dataset and its owner, available licenses, pricing, when relevant, and the dataset's range, quality, and domain.

A step further, such a system would offer license brokering and management to data owners and data consumers, including an easy to use, visual expert system ("wizard") that helps data owners to find the best possible license model custom-tailored to their needs and wishes. A dataset owner could then determine in which way the dataset can be used: for research, commercial or non-commercial use, whether anonymization or pseudonymization is required, whether only summaries can be published and in what aggregation, what fees and what terms of non-disclosure apply.

To further facilitate data reuse, such a system would enable data owners to enforce fee collection for various license types, for commercial or non-commercial use. This way of monetizing available data entices further collaboration and data sharing between research groups.

For commonly used file and stream formats, the platform could offer value-added services to data owners that enrich datasets and simplify data preparation, including auto-anonymization or pseudonymization, geo-coding, geo-tagging, visualization, auto-aggregation, and the (semi-) automatic generation of sample data.

Such a system would include, for each deposited dataset, acknowledgments to sponsoring agencies and programs, and references to research papers and patents resulting from data access. The system would generate reports on how many papers, patents, projects, and other artifacts result from data access, sorted by dataset, owner, or sponsoring agency or program. In combination with the rating of datasets and transactions by both owners and users, this reporting generates a certain degree of peer pressure, in addition to the formality of the license terms, to ensure full and proper acknowledgment.

2.2 Services to Data Users

With the envisioned platform, interested researchers would be able to browse deposited datasets by category, domain, license, collection and deposition date, and other attributes, download sample data, and check available licenses. For common file and stream formats, the system would offer data previews in a web browser, including table-based views, aggregations, simple statistics, and visualization. If questions remain, the platform allows the interested data consumer to contact the dataset owner. It is feasible that a repository system could interface, should the collection of a fee be required, to secure external payment services to facilitate the transaction.

Once a deal is struck, the user can download it or use the system's API to access the dataset.

To encourage data sharing and cooperative behavior, a system might offer trust-building community functions, including the rating of datasets and transactions, as well as integration with research social networks. Using gamification measures, users could be encouraged to review datasets and to rate them according to their quality and range. In addition, the platform could help to build communities of users curating groups of datasets, and to discuss their strengths and weaknesses.

Such a system would offer advanced functions for data access, programmed searches, and allow data owners to dynamically update datasets once new data becomes available – data consumers' listeners registered with corresponding datasets would be automatically notified as soon as any updates take place. The system would also allow code and data integration, would provide an interface to and manage references to Git code repositories. Optionally, the data resulting from these algorithms could in turn be stored in the system via its API, thereby adding value to already present data.

The envisoned system would be open to all computer science researcher, regardless of specialization and research field, whether working in industry or in academia. While other repositories focus on bringing together specialized data based on distinct fields of origin (e.g. geology, genetics, or marine biology), the platform would focus on data supporting the computing research data consumption needs, regardless of the domain of data origin. It would therefore especially suit the needs of modelling and simulation researchers.

Figure 1 depicts a distillation of the envisioned services.

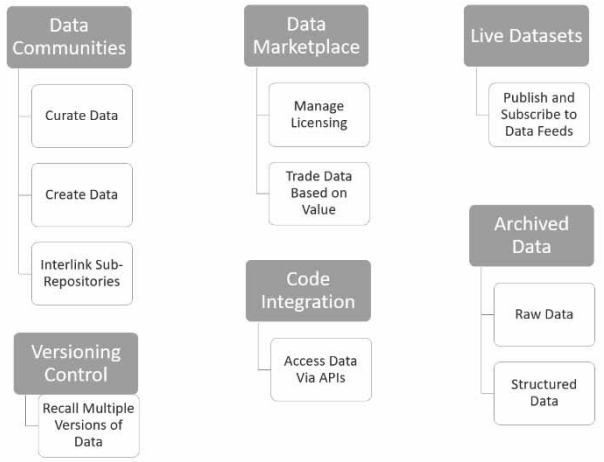

Figure 1: A summary of envisioned functionality.

3 Repositories for Simulation Data

3.1 Overview

The multitude of 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.

An overwhelming majority of these repositories are designed to mimic digital libraries, which have data storage and retrieval features, might support basic versioning, and contain multidisciplinary datasets. These digital libraries have basic functionality and do not support advanced data retrieval via an API, or the interlinking of related datasets. Out of the 19 repositories examined as a sample (see Figure 2), only six of the systems reviewed that provide more advanced functionality supporting the requirements of computer scientists, more specifically to simulation modellers (see Figure 3).

These repositories with advanced functionality beyond that of a standard digital library are Figshare [8], Zenodo [23], Unidata Internet Data Distribution [22], CKAN @ IoT Lab [5], CITK [4], and GitHub [9].

3.2 A Closer Look

The following six dataset repositories contain features especially geared towards supporting computer science research (see [17] and [16]):

Figshare. Figshare is general-purpose cross-disciplinary data repository based in Great Britain. It is one of the most popular repositories and houses more than 500,000 datasets, which is more than any other data repository surveyed. It has a seamless intuitive user interface and allows for storage, sharing, interlinking, and discovery of a multitude of artifacts, including figures, media, datasets, file-sets, posters, papers, presentations, thesis, and code. Figshare supports access via a REST API as well as the industry standard OAI-PMH.

User workspace offered by Figshare allows users to create and manage projects by adding or removing artifacts related to it. It has facilities to link collections to

the project and observe the project activity over time. The workspace is another key differentiating feature that motivates the data publishers and consumers to work within the service's ecosystem and perpetuate it by storing and sharing the results of their research based on data gathered from Figshare through the service itself.

In-browser data preview is a data visualization feature of Figshare. It allows for data of various types to be visualized instantly within browsers to provide the user with a general sense of the data before the user initiates a download. This functionality reduces the stress on the overall system by helping to prevent the user from downloading unneeded datasets.

Figshare is missing the support for live updates to datasets, data versioning mechanisms, community-based data curation, and only allows to link code repositories from GitHub. The service does not host code repositories itself, which limits the level of integration between data and code repositories. Furthermore, it does not provide means to trade data.

Zenodo. Zenodo is an open data catch-all cross-disciplinary repository for research funded by the European Commission, also supported and funded by CERN. Similarly to Figshare, the service provides access via REST and OAI-PMH. Each upload gets its own Digital Object Identifier (DOI), which makes it uniquely identifiable and citable. The repository accepts various data types, including publications, datasets, software, and presentation. It also allows for in-browser data visualization, which enables users to judge the fit of data for a particular purpose prior to download initiation. The service provides the facilities to identify grants used in research as well as flexible licensing that allows for sharing of datasets among communities.

Zenodo's approach to the user workspace within a data repository service is somewhat different from that of Figshare. The service is organized around the concept of communities. Consumers and publishers are encouraged to organize into communities, where they can create their own sub-repositories and curate the data that gets deposited. It means that Zenodo has the facilities to self-organize into meaningful groups that work on similar goals and datasets as well as self-curate to ensure that data is relevant, valid, and generally useful.

Name	Key Functionality	Reference
BABS	Basic digital library for the humanities	[2]
CCITK	Storage of data, data processing code, and derived data sets	[4]
CIESIN	Basic digital library for earth science data	[3]
CKAN	Storage of data, data processing code, and derived data sets	[5]
Clarin	Basic digital library for language resources	[6]
Dataverse	Repository system for basic digital libraries	[7]
Figshare	User orientation, in-browser visualization, data versioning	[8]
GEON	Basic digital library	[10]
GitHub	Storage for code and accompanying data, versioning	[9]
NatureServe	Digital repository for biodiversity data, provides API for access	[11]
OLAC	Basic digital library for language resources	[12]
Pachyderm	Platform to host digital libraries with additional versioning	[13]
PredictDB	Digital library for genome data and prediction models	[14]
RAMADDA	Digital library for satellite data	[15]
SNAP	Repository for network-related data from several disciplines	[18]
UA-CR	Basic digital library	[19]
UCI	Basic digital library for large standardized data sets to test machine learning algorithms	[20][21]
Unidata	Digital library plus real-time delivery	[22]
Zenodo	Community orientation, in-browser visualization, data versioning, high searchability	[23]

Figure 2: A sample of 19 data repository systems was examined.

The service is implemented using Invenio, a free open-source digital library framework originally developed at CERN. The framework covers all aspects of digital library management and allows for diverse content types, including articles, books, journals, photos, videos, and datasets. It provides digital library features, including navigable collection tree, powerful search engine, flexible metadata, collaborative features, and personalization.

With this type of framework, it is relatively inexpensive to get a basic digital library-type data repository running. Consequently, a software engineering team may quickly shift focus towards integration of more advanced functionality.

Much like Figshare, Zenodo is missing the means to trade data, live dataset updates, data versioning mechanisms, and data market. Unlike Figshare, it does provide the support for content curated via communities.

Unidata. Unidata Internet Data Distribution (UIDD) is a community of 260 universities sharing tools to disseminate near real-time earth observation data online. While offering the standard data storage, retrieval and discovery features, this service is designed to automatically deliver certain datasets to subscribers as soon as the data becomes available. In other words, a publisher of data can establish a link to the repository to deposit data from the publisher's sensors in real-time. Subsequently, a consumer of data can subscribe to the published data feed and receive dataset updates in near real-time. This concept of "live datasets" is appealing as it makes the datasets dynamic, which enables not only research opportunities based on most up-to-date data, but also to some extent the creation of applications that showcase whether or not the research findings remain valid when provided with new data points, which were not in the original data set.

Unidata is missing all of the advanced functionality of Figshare and Zenodo. The only functionality that sets it apart from a digital library is "live datasets", which is surprisingly absent from all other repositories with advanced functionality.

CKAN @ IoT Lab and **CITK** are both data repository platforms that focus on research data and software code integration. These repositories position themselves as toolkits for researchers, because they implement facilities to store and manage datasets as well as software code associated with datasets.

This type of functionality is benefitial for researchers in the computer science field because many researchers use custom-built software to perform research using external datasets, and it only makes sense to be able to store the code alongside the corresponding data set. Furthermore, it makes sense to store the resulting dataset as a derivative of the original and interlink the original dataset, software code, and result dataset within the same repository. This idea builds further on traceability of the evolution of datasets and significantly improves the motivational aspect for the data consumers to store and share their work within the repository ecosystem.

CKAN also functions as an open source data portal. The implemented features allow to publish and find datasets, store and manage data, and engage with users. It is also highly extendable and customizable, has advanced geospatial and visualization features, and includes a RESTful JSON API for querying and accessing the dataset information. This software can be used to create a data repository with basic features quickly and build extensions necessary to support more advanced features tailored to the computer science community.

CKAN and CITK 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 data exchange markets.

GitHub. At the opposite end of the dataset/code spectrum is the widely-used GitHub code repository. The repository is built using Git open-source software, which is created to share, track, manage and execute simple and complex software projects. Git is one of the most widely used team software code management technologies by computer scientists around the world. It allows to create a code repository, share it, and collaborate on it while mitigating conflicts between the changes made to the code by the participants in the sharing process. It also allows for the creation of derivations of the source code, thereby enabling project evolution, while maintaining full traceability of changes made by all participants. GitHub is not designed for dataset storage, the underlying Git technology was created with source code in mind. However, the overall code storage, sharing, management, evolution, and traceability principles are applicable to pure dataset repository realm.

Since code is essentially data, a data repository geared towards researchers within the computer science field should provide the facilities to deposit and manage code alongside the relevant datasets.

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.

Figure 3 depicts an overview of the main function groups aimed at supporting computer scientists that the six discussed data repositories feature.

Feature	CITK	CKAN	Figshare	GitHub	Unidata	Zenodo
Communities	×	×	×	×	×	✓
Marketplace	×	×	×	×	×	×
Code Integration	✓	✓	✓	✓	✓	✓
Versioning	×	×	×	✓	×	×
Live Datasets	×	×	×	×	✓	×

Figure 3: An overview of the main feature groups of selected data repository platforms aimed at supporting computer scientists.

4 Conclusion

This paper presented an overview of 19 data repository systems in the area of computer science. Out of these reviewed systems, only six have advanced feature sets that go significantly beyond standard digital libraries. No single data repository provides a combination of features and tools geared towards simulation projects conducted at smaller laboratories, and none offers features that would allow for purchase or sale of data.

Among other considerations, the existing platforms are especially failing to create a marketplace environment where computer scientists are enticed 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 scripts that derive versions of originals, which can be used for further, non-obvious modeling and analysis, while appropriately crediting the original dataset.

References

[1] Assante M, Candela L, Castelli D, Tani A. Are Scientific Data Repositories Coping with Research Data Publishing? *Data Science Journal*, vol. 15, 2016.

[2] "Bayerischen Staatsbibliothek und Staatsarchive, Langzeitarchivierung", https://www.digitale-sammlungen.de/index.html, accessed 17-Feb-2020.

[3] "Center for International Earth Science Information Network", http://www.ciesin.org/data.html, accessed 17-Feb-2020.

[4] "Cognitive Interaction Toolkit", https://toolkit.cit-ec.uni-bielefeld.de, accessed 7-Jan-2017.

[5] "CKAN @ IoT Lab", http://ckan.iotlab.eu, accessed 7-Jan-2017.

[6] "CLARIN". https://www.clarin.eu/, accessed 17-Feb-2020.

[7] "Dataverse", https://dataverse.org/, accessed 17-Feb-2020.

[8] "figshare", https://figshare.com, accessed: 7-Jan-2017.

[9] "GitHub", https://github.com, accessed 7-Jan-2017.

[10] "GEONGRID – Global Earth Observation Network", http://geongrid.org, accessed 16-Dec-2018.

[11] "NatureServe Expolorer", http://explorer.natureserve.org/, accessed 18-Feb-2020.

[12] "OLAC – Open Language Archives Community", http://www.language-archives.org/, accessed 17-Feb-2020.

[13] "Pachyderm", https://www.pachyderm.com/, accessed 17-Feb-2020.

[14] "PredictDB Data Repository", http://predictdb.org/, accessed 18-Feb-2020.

[15] "RAMADDA on the NSF Jetstream Cloud", https://ramadda.scigw.unidata.ucar.edu, accessed 18-Feb-2020.

[16] Schomm F, Stahl F, and Vossen G. Marketplaces for data. *ACM SIGMOD Record*, vol. 42, no. 1, p. 15, 2013.

[17] Stahl F, Schomm F, Vossen G, and Vomfell L. A classification framework for data marketplaces. *Vietnam Journal of Computer Science*, vol. 3, no. 3, pp. 137–143, 2016.

[18] "SNAP – Stanford Large Network Dataset Collection", https://snap.stanford.edu/data/, accessed 18-Feb-2020.

[19] "University of Arizona – Campus Repository", https://repository.arizona.edu/arizona/, accessed 18-Feb-2020.

[20] "UCI Knowledge Discovery in Databases Archive", http://kdd.ics.uci.edu/, accessed 18-Feb-2020.

[21] "UCI Repository", http://ics.uci.edu/~mlearn/MLRepository.html, accessed 7-May-2017.

[22] "Unidata" Internet Data Distribution, http://www.unidata.ucar.edu/projects/index.html#idd, accessed 7-Jan-2017.

[23] "Zenodo – Research. Shared.", https://www.zenodo.org, accessed 7-Jan-2017.