Big Data Bags: A Scalable Packaging Format for Science

Mike D'Arcy*, Kyle Chard†, Ian Foster†, Carl Kesselman*, Ravi Madduri†, Nickolaus Saint†, Rick Wagner†
†*The University of Chicago, Chicago IL, USA*
**Information Sciences Institute, University of Southern California, Marina del Rey CA, USA*

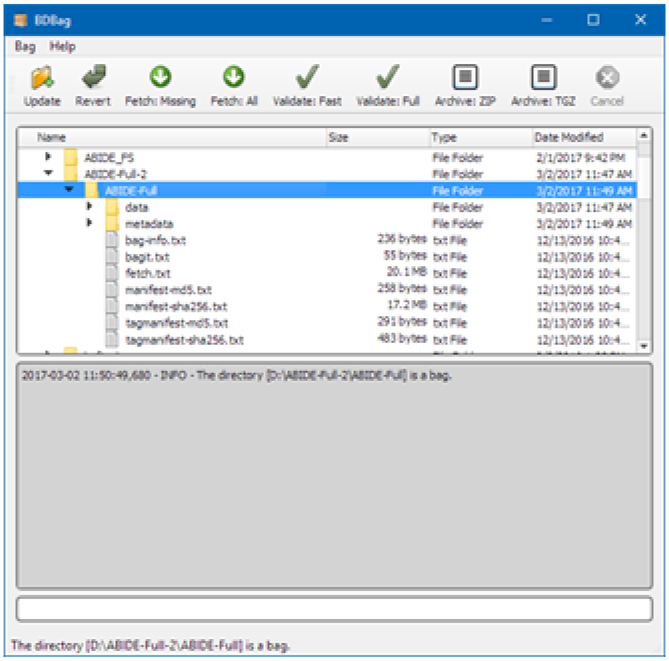
I. Introduction

The need to describe and exchange large and complex data underlies the vast majority of science conducted today. Such needs arise when downloading data from a repository, moving data between remote locations, exchanging data between collaborators, and even publishing data as part of the publication process. While such examples are common, it is surprisingly difficult to describe and exchange data, and it is even more difficult when datasets are large and span multiple storage locations. To address some of these challenges we proposed the Big Data Bag (BDBag) [1] as a data packaging format for representing and describing complex, distributed, and large datasets. In this presentation, we outline the BDBag model and describe three scenarios in which it is currently being used.

BDBag is designed to provide a simple and convenient way of defining and describing the contents of a dataset. BDBags extend the BagIt specification [2], using it to provide basic metadata, enumerate the contents of a dataset, and as a standard packaging format for exchange. The BagIt specification specifies how data files are hierarchically named in a directory structure, includes a manifest of all of the data objects including checksums on their contents, and specifies that bags can be serialized into a ZIP file. One of the most significant advantages of the bagIt specification is that it allows files to be either included in the bag or to be directly referenced via a URL in the `fetch.txt` file. This allows for the exchange of 'holey' bags which need not contain a copy of all data resources and thus can be significantly smaller than if all files were to be included.

The BagIt specification provides little guidance on how metadata should be encoded in a bag. To address the need to describe bag contents, we developed the BDBag BagIt profile [3] which specifies the use of the Research Object (RO) framework [4] for encoding metadata. The RO format allows for the description of the resources, attribution, provenance, and structured and unstructured annotations associated with resources in the bag. The BDBag BagIt profile requires that compliant bags include a RO `manifest.json` file that outlines for each resource (i.e., file in the data directory or remote files) the name, type, and annotations associated with that resource. Thus, BDBag users are able to specify metadata and even relationships between resources in the bag in a simple and structured manner.

To simplify use of BDBags we have created both a command line client and a graphical user interface (GUI). The BDBag command line client enables creation, validation, (de)serialization (via a ZIP file), and downloading of remote files, all through a command line interface. The BDBag GUI provides a graphical user interface for working with BDBags. Users can create and update bags, and also validate, archive, and fetch remote files. The GUI is shown in Figure 1. The BDBag client and GUI are able to automatically resolve references to remote files and to download files via Globus [5], HTTP, or FTP.



Name	Size	Type	Date Modified
ABIDE_FS		File Folder	2/1/2017 9:42 PM
ABIDE-Full-2		File Folder	3/2/2017 11:47 AM
ABIDE-Full		File Folder	3/2/2017 11:49 AM
data		File Folder	3/2/2017 11:47 AM
metadata		File Folder	3/2/2017 11:49 AM
bag-info.txt	236 bytes	txt File	12/13/2016 10:4...
bagit.txt	55 bytes	txt File	12/13/2016 10:4...
fetch.txt	20.1 MB	txt File	12/13/2016 10:4...
manifest-md5.txt	258 bytes	txt File	12/13/2016 10:4...
manifest-sha256.txt	17.2 MB	txt File	12/13/2016 10:4...
tagmanifest-md5.txt	291 bytes	txt File	12/13/2016 10:4...
tagmanifest-sha256.txt	483 bytes	txt File	12/13/2016 10:4...

```
2017-03-02 11:50:49,680 - INFO - The directory [D:\ABIDE-Full-2\ABIDE-Full] is a bag.
```

The directory [D:\ABIDE-Full-2\ABIDE-Full] is a bag.

Figure 1: BDBag graphical user interface.

II. Use cases

We have used BDBags as a common format for data validation, tracking aggregations, referencing remote files, and capturing metadata and provenance on several projects. Here we outline their use in three diverse scenarios: scientific reproducibility, exchange of biomedical data, and for exporting virtual cohorts from biomedical catalogs. Several other scenarios have been outlined in our prior work [1].

In fields such as biomedicine there is a growing reproducibility problem caused in part by increasing data volumes and complexity and also due to the inherent distribution and complexity of analysis processes [6]. In order to mitigate these challenges, there is a need for easy-to-use packaging formats that enumerate dataset contents, scale to huge dataset sizes, include descriptive metadata, provide methods for tracking provenance, and enable seamless exchange. Based on these requirements, we adopted BDBags as the cornerstone of a reproducible big data pipeline that was used to construct genome-wide maps of candidate transcription factor binding sites via footprinting methods [7]. The pipeline contains a number of disjoint steps including download of huge datasets from a public repository, various computational processing pipelines in Galaxy on the cloud, and application of ad hoc processing steps by researchers on local computers. We showed, via a user study, that the use of BDBags and other methods enabled a diverse group of participants to reproduce a complex biomedical analysis [8].

We have integrated BDBags as the preferred export format for data collections in the DERIVA platform for scientific asset management [9]. In this use case, an export service takes a description of a set of data files and associated descriptive metadata (e.g. a set of image files and attributes describing the sample being imaged) and serializes it into a BDBag. Typically, a tabular representation of the image metadata would be directly included in the bag as a comma separated value (CSV) file, while the larger image data would be referenced remotely. The remote images within the data directory would be logically organized according to a sample identifier (which might be one of the values contained in the metadata table), although other organizations, such as by imaging type might be used depending on the nature of the collection. DERIVA can be configured to download the BDBag directly to a client computer, or to put the bag into a shared cloud based storage system, such as S3, and return a permanent identifier to the bag to enable unique reference and sharing.

As part of a pilot project to enable interoperation and federated access across several biomedical data providers we faced the need for a uniform format for exchanging data and metadata between data repositories, cloud-based analysis platforms, and other services. To address this need we used BDBags to package local files and references to remote data as well as metadata describing those data using a project-specific metadata model. The use of BDBag enabled a single interface across the set of distributed services, scalability to huge datasets via data references, inclusion of metadata using different metadata schema, and validation of the integrity of the local, referenced, or metadata files included within the bag. BDBag also allowed for the capture of provenance associated with analyses by storing the output of an analysis as data in the bag and including references to the workflow, its parameters, and the input data.

III. Summary

BDBags provide a simple yet powerful way of specifying, sharing, and managing complex, distributed, and large datasets. BDBags rely on BagIt for enumerating dataset contents, Research Objects for describing a dataset using arbitrary metadata, and custom tools for integration in arbitrary research processes.

BDBag source code and tooling is openly available on GitHub (https://github.com/fair-research/bdbag)

References

[1] K. Chard *et al.*, "I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets," in *IEEE International Conference on Big Data (Big Data)*, Dec 2016, pp. 319–328.

[2] J. Kunze *et al.*, "The BagIt file packaging format (V0.97)," Internet Engineering Task Force, Internet Draft (work in progress), draft-kunze-bagit-11.txt, Tech. Rep., 2015.

[3] S. Soiland-Reyes, "Research Object BagIt archive," researchobject.org, https://w3id.org/ro/bagit. Visited July 1, 2019.

[4] S. Bechhofer *et al.*, "Research Objects: Towards exchange and reuse of digital knowledge," in *Workshop on The Future of the Web for Collaborative Science*, 2010, available from Nature Precedings http://dx.doi.org/10.1038/npre.2010.4626.1.

[5] K. Chard, S. Tuecke, and I. Foster, "Efficient and secure transfer, synchronization, and sharing of big data," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 46–55, Sep. 2014.

[6] A. W. Toga *et al.*, "Big biomedical data as the key resource for discovery science," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1126–1131, 2015.

[7] C. C. Funk *et al.*, "Atlas of transcription factor binding sites from encode dnase hypersensitivity data across 27 tissue types," *bioRxiv*, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/01/27/252023

[8] R. Madduri *et al.*, "Reproducible big data science: A case study in continuous fairness," *PLOS ONE*, vol. 14, no. 4, pp. 1–22, 04 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0213013

[9] A. Bugacov *et al.*, "Experiences with deriva: An asset management platform for accelerating escience," in *13th IEEE International Conference on e-Science (e-Science)*, Oct 2017, pp. 79–88.