HORIZON 2020 FRAMEWORK PROGRAMME

IOStack

(H2020-ICT-2014-7-1)

Software-Defined Storage for Big Data on top of the OpenStack platform

D2.1 Data Management Plan

Due date of deliverable: 30-06-2015
Actual submission date: 30-06-2015

Start date of project: 01-01-2015 Duration: 36 months

Summary of the document

Document Type	Deliverable
Dissemination level	Public
State	v1.0
Number of pages	9
WP/Task related to this document	WP2 / T2.1
WP/Task responsible	URV
Author(s)	Raúl Gracia-Tinedo
Partner(s) Contributing	URV
Document ID	IOSTACK_D2.1_Public.pdf
Abstract	The present document provides the data management plan of the IOStack project, in particular for the management of open data. It describes the overall open data strategy in IOStack as well as the concrete actions that the consortium will undertake to transform the resulting project data assets (publications, datasets and source code) into open data.
Keywords	Data Management, Open Data, Science 2.0

Table of Contents

1 Executive summary

Open data is becoming increasingly important for maximizing the excellence and growth of the research activity in Europe. In this sense, the motivation of the IOStack project is very aligned with the foundations of open data: IOStack aims at building a Software-Defined Storage toolkit on top of OpenStack, which is the largest open source project in cloud technologies. Thus, our next step is to plan how this natural synergy between IOStack and the open source community is materialized into a set of open data assets available to the general public.

The present document provides the data management plan of the IOStack project, in particular for the management of open data. It describes the overall open data strategy in IOStack as well as the concrete actions that the consortium will undertake to transform the resulting project data assets into open data. Essentially, these actions are directed to the three main data assets produced in IOStack: publications, datasets and software source code. We also describe how we will make use of dissemination mechanisms to increase the impact and visibility of the open data generated in IOStack.

2 Open Data in IOStack

Open data is becoming increasingly important for maximizing the excellence and growth of the research activity in Europe. Since the beginning of the 2000's, Europe is leading a major initiative to make publicly founded research projects *actually public* by taking into special consideration the *openness and transparency* of the management of a project's results. And no need not mention, most of these results and research assets can be classified as *data* (e.g., research papers, datasets). Clearly, the wave of open data continues and it is being strengthened in the H2020 framework.

The value of open data is clear: it improves circulation, access to and transfer of scientific knowledge and tools, which in turn, optimizes the impact of publicly-funded scientific research. In this sense, the motivation of the IOStack project is very aligned with the foundations of open data. Actually, IOStack aims at building a Software-Defined Storage toolkit on top of OpenStack, which is the largest open source project in cloud technologies. This level of commitment with an open source community gives a sense on the open data strategy of the project as a whole.

It is worth mentioning that some partners of the IOStack consortium are already adhered to open data standards in other EU-funded projects. For instance, URV —coordinator of the IOStack project— is currently implementing *green* open data policies for datasets, research papers and software source code in the context of the FP7 CloudSpaces project[1]. Thus, our experience on opening prior research results guarantees an effective application of open data policies in IOStack as well.

In this document, we define the policies and mechanisms that will help us to transform the project's outputs and research results into open data.

3 Types of Data Generated/Collected in IOStack

In this project, we consider 3 main sources of assets that can be subject of open data policies: *research papers*, *datasets* and *software source code*.

Research papers: In IOStack, research papers are the main vector of propagating our research contributions to the appropriate audience —both conferences and journals. During the project, we will target high-quality publications in order to maximize the impact of our research discoveries. In any case, as we detail later on, all the publications related to the IOStack project will be made publicly available following *green* open data standards.

Datasets: Often, a research publication is based on or has as a result a dataset. Datasets may contain any type of information that can lead to reproduce or verify the claims supported in the publication itself. In IOStack, we foresee the generation of various datasets ranging from company use cases workloads to data capturing the performance results of benchmarking our SDS toolkit. Such information will be of great interest for the community in order to foster research in this field. Datasets will be also made publicly available in conjunction with the necessary metadata and tools for processing the dataset.

Software source code: The ultimate objective of IOStack is to build an open-source SDS toolkit for OpenStack. From an engineering perspective, such an ambitious goal cannot be achieved as a single, monolithic piece of software, but rather as a set of advanced software components converging on a single architecture. Our objective is to achieve both a proper software management in IOStack and transform the source code into open data from the very beginning. As we detail next, all the partners are contributing to a public and centralized code management system. This makes the development of the project open and transparent for the public.

In what follows, we depict a battery of actions to convert the previous three types of data assets into open data.

4 IOStack Open Data Policies and Standards

Next, we aim at describing the overall open data strategy of IOStack as well as the concrete measures to make data assets publicly available (see Fig. 1).

[1] http://cloudspaces.eu

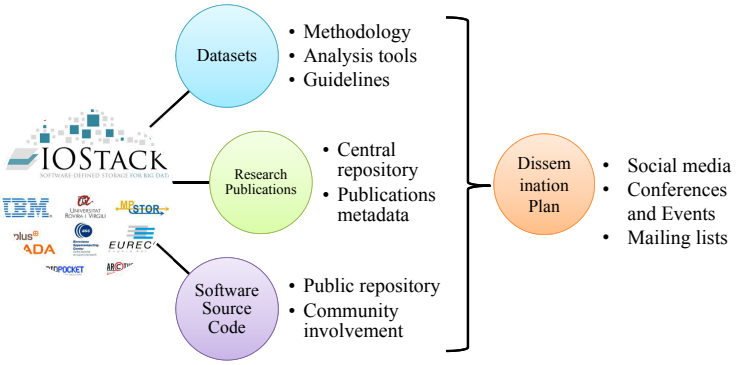

Figure 1: High-level open data strategy in IOStack.

As can be observed in Fig. 1, the open data strategy of IOStack is based on 4 main action lists; one action list for each type of data asset produced during the project and the dissemination plan, which is a particular action list to promote the impact of the open data produced in IOStack. Thus, IOStack implements an integral plan for generating open data and promoting it in order to achieve the widest dissemination possible.

All the elements in Fig. 1 have a common denominator: the IOStack web site[2]. The IOStack is being actively maintained and offers easy access to all the data assets of the project (publications, code, datasets), the project's deliverables and the social media accounts of the project[3].

We continue by depicting the different action lists of the IOStack open data strategy.

4.1 Research Papers

Before going any further, we should consider that there are two main approaches to implement open data on research papers: *gold* and *green* open data [1]. In the former case, researchers can publish in an Open Access (OA) journal, where the publisher of a scholarly journal provides free online access. On the latter case, researchers can deposit a version of their published works into a subject-based or institutional repository.

Although the *gold* open data approach has gained strength in the latest years, *we advocate for the green approach* due to a strong reason: Today, most high-impact conferences and journals are not yet Open Access (OA). Consequently, adopting a pure gold open data approach may be in detriment of the potential impact of IOStack publications. For this reason we adopt a green open data approach in IOStack.

In what follows, we describe an action list to enable better access to the scientific publications of IOStack in order to convert them into *green* open data.

- **Self-archiving**: Self-archiving is considered a valid route to make a research paper open data (green). URV has created a repository to archive all the publications related to the project. Concretely, the repository for publications is embedded into the IOStack official web site and can be accessed at "http://iostack.eu/publications". The repository offers a user-friendly interface that permits to navigate across multiple publications.

- **Deposit procedure**: In each publication entry in the repository, we deposit a machine-readable copy (e.g., PDF) of the final version or final peer-reviewed manuscript accepted for publication. We will attempt to deposit the final version of the manuscript as soon as possible, trying to avoid any embargo period.

- **Durability and availability**: Internally, the server that hosts the publication repository —and the IOStack web site— integrates disk-level redundancy to support failures and data corrup-

[2]http://iostack.eu
[3]https://twitter.com/iostackproject

tion. Moreover, URV backs-up the information of that server the every week in other machines. To maximize the durability and availability of open access to our research publications, each partner will self-archive its own publications so in case of catastrophic events in URV's infrastructure publications can still be available. For example, URV already keeps two separate repositories for the publications of IOStack and the publications of Arquitectures i Serveis Telemàtics (AST) research group ("http://ast-deim.urv.cat/web/publications").

- **Publication metadata**: Every paper in the IOStack publication repository contains the associated metadata that describes the type and topic of the publication (abstract), as well as the original publisher, venue and Document Object Identifier (DOI).

- **Standard methodologies**: Apart from the way research publications are made publicly available to users, we believe that it is also important to implement standard and open methodologies during the elaboration of research articles. To this end, as a part of the benchmarking framework of IOStack (d.2.2), the consortium will resort to exiting open benchmarks and datasets in order to validate research contributions.

With this initial battery of actions, we aim at transforming research papers of IOStack into green open data easily accessible by the general public.

4.2 Datasets

In many cases, a research publication has associated a dataset, either as a source of information to extract novel observations or as a result of the research process. Our aim is to deposit at the same time the research data needed to validate the results of the associated research publications. Next, we specify the action list that we undertake to implement green open data policies on datasets:

- **Self-archiving**: Similarly to the approach adopted for research publications, URV has created a repository to store all the datasets related to IOStack. To ease the location of datasets, the repository for datasets is also embedded into the IOStack official web site and can be accessed at "http://iostack.eu/datasets".

- **Durability and availability**: In the general case, the procedure to maintain the availability and durability of datasets is the same as explained for research publications, since all these data assets reside in the same physical servers. However, a distinguishing point for datasets is that we also make active use of them in a data processing cluster located at AST research group labs (URV). Internally, this cluster implements 3-way replication, so datasets have an additional physical infrastructure to maximize their durability in case of damage of the servers dedicated to host the IOStack web site and the publications.

- **Open formats and metadata**: Datasets will be generated making use of open formats instead of proprietary ones (e.g., Microsoft Excel). Concretely, we expect to make extensive use of the Comma Separated Value (CSV) format, which is generic enough to express very different types of information. Of course, for every dataset in the repository, we will provide the required dataset metadata to understand the *topic*, *purpose*, *collection/generation methodology* as well as an explanation of the different *fields* of the dataset. This will improve the researchers' accessibility to the datasets generated in IOStack.

- **Parsing tools**: Sometimes, it is necessary to parse datasets to easily extract particular parts of the information contained inside it. If parsing tools are necessary for the correct analysis of the datasets, we will provide the tools jointly with the dataset in the repository.

With this action list on datasets we will facilitate their open access to convert them into green open data.

4.3 Source Code

The ultimate objective of IOStack is to provide a SDS toolkit —i.e., software source code— on top of the OpenStack platform. This means that during the development of the project, we should adopt open data policies from the very beginning regarding the produced source code. Our strategy will not only leverage the results at the end of the project as open data, but it also makes the source code as open from the entire software life-cycle. In turn, this paves the way for the involvement of the OpenStack community in IOStack as well.

- **Central code repository**: To make the source code open to the general public, we created a code repository in GitHub for IOStack at "https://github.com/iostackproject". This repository has been also linked to the IOStack web site ("http://iostack.eu/software"). GitHub is currently one of the most popular code management systems due to the advanced features and easy management that it provides to developers. This has various potential benefits to the management and dissemination of IOStack source code: for intance, GitHub is well-known across developer communities, which facilitates the access to the source code of IOStack. Moreover, GitHub offers a plenty of options to fork/branch/merge versions of a software project that enables third-parties to easily extend the source code developed in IOStack (even for internal use).

- **Availability and durability**: GitHub is a cloud-based system. This means that, internally, the code repositories in GitHub are stored across several physical machines, even in distinct geographical regions. Therefore, the availability and durability of IOStack source code is ensured and delegated to GitHub, conversely to our self-archiving approach for research papers and datasets.

- **Licensing**: Whenever possible, we will retain the copyright and grant adequate licenses to the source code created in IOStack. In general, the code will be released under open licenses such as Apache License 2.0 or GNU General Public License 3.0. Broadly speaking, these licenses provide the user of the software the freedom to use the software for any purpose, to distribute it, to modify it, and to distribute modified versions of the software, under the terms of the license, without concern for royalties [2]. However, the intellectual property of the source code is kept: For instance, the Apache License requires preservation of the copyright notice and disclaimer, which are related to the project [3].

- **Source code metadata and "how to"**: As a standard practice in the open source community, every software project in the IOStack repository will include a README file to help the user in the installation process, testing and first steps using the software. This will ease the use and adoption of the source code produced in IOStack.

This action list will make IOStack source code as open data assets. Next, we explain how we will maximize the impact of open data in IOStack through our Dissemination plan.

4.4 Disseminating Open Data

From a practical viewpoint, the generation of open data is only a part of the work that should be done in IOStack regarding data management. We believe that the dissemination of open data results is as important as its generation, since a hidden open data item might be useless for the general public.

In the following, we depict the major actions related to dissemination of open data assets in IOStack:

- **Dataset link in research publications**: In IOStack, all the research publications that make use of datasets generated within the project should cite the repository where the datasets live. Although this is a common practice in the research community, it is important to remark that research publications may have high visibility that can be beneficial for the dissemination of open data in IOStack.

- **Promotion in conferences/events**: One great benefit of publishing in high-impact conferences is that one has direct access to the research community during the presentation of a paper. For this reason, if a research publication involves producing/collecting datasets, the responsible IOStack partner will disseminate not only the particular research contributions of the publication but the datasets as well.
- **Social media and mailing lists**: IOStack already has a Twitter account to disseminate the news and events related to the project ("http://twitter.com/iostackproject"). The consortium will use this account collaboratively to amplify the dissemination of our open data contributions. Moreover, MPStor —leader of the dissemination plan— will make use of large mailing lists to also notify industrial/research organizations about the advances and contributions of the project in this sense.
- **Internal reutilization**: The IOStack consortium will maximize collaboration between partners to exploit the open data generated during the project. In addition to save up unnecessary efforts, we will act as testbeds of our own open data. This will lead to future enhancements of the action lists defined in this document.

5 IOStack Open Data Assets

In this section, we enumerate the available data assets in the IOStack project to date. Of course, this is only an snapshot of the current state of IOStack's results. As the project advances, we will keep the open data assets of the project updated in the data management plan.

5.1 Research Papers

SDGen: Mimicking Datasets for Content Generation in Storage Benchmarks

- *Author Partners*: URV and IBM.
- *Published at*: 13th USENIX Conference on File and Storage Technologies (FAST'15). February 16-19, 2015, Santa Clara, CA, USA.
- *Deposit format*: PDF File.
- *Available at*: http://iostack.eu/publications/download/publications/2-fast-sdgen
  http://ast-deim.urv.cat/web/publications?view=publication&task=show&id=558
  https://www.usenix.org/conference/fast15/technical-sessions/presentation/gracia-tinedo
- *Archiving and preservation*: The publication is freely available and archived at IOStack repository, AST lab servers and USENIX Association.
- *Type of open data*: Green open data.

5.2 Datasets

There are datasets in collection/generation phase, but they are not available yet.

5.3 Source Code

Storlets

- *Responsible Partner*: IBM.
- *Software description*: The Storlet project provides computation-close-to-data functionalities to the IOStack architecture for object storage.
- *Available at*: https://github.com/iostackproject/swift-storlets.

- *Archiving and preservation*: Freely available and archived at GitHub.
- *License*: Apache License 2.0.
- *Status*: Development.

SDS Controller for Object Storage

- *Responsible Partner*: URV, IBM and BSC.
- *Software description*: The SDS Controller for object storag will provide unified management, orchestration and automation of the services that form the IOStack toolkit.
- *Available at*: https://github.com/iostackproject/SDS-Controller-for-Object-Storage.
- *Archiving and preservation*: Freely available and archived at GitHub.
- *License*: Apache License 2.0.
- *Status*: Development.

IO Bandwidth differentiation

- *Responsible Partner*: BSC.
- *Software description*: The Bandwidth Differentiation service will enable the IOStack toolkit to regulate the bandwidth assigned to each tenant in a multi-tenant analytics platform on Swift.
- *Available at*: https://github.com/iostackproject/IO-Bandwidth-Differentiation and https://github.com/iostackproject/IO-Bandwidth-Differentiation-Client.
- *Archiving and preservation*: Freely available and archived at GitHub.
- *License*: Apache License 2.0.
- *Status*: Development.

SDGen

- *Responsible Partner*: URV and IBM.
- *Software description*: SDGen is a synthetic data generator that can emulate the compression properties of real datasets, which is a fundamental aspects when it comes to benchmark data reduction techniques in IOStack.
- *Available at*: https://github.com/iostackproject/SDGen.
- *Archiving and preservation*: Freely available and archived at GitHub.
- *License*: GNU General License 3.0.
- *Status*: Released.

6 Final Remarks

Nowadays, open data is becoming a key enabler for the Europe Research Area in order to maximize the impact and profit of publicly funded research. In this document, we described the strategy and actions that we are undertaking in IOStack for transforming the data assets of the project (datasets, publications, source code) into open data. Our objective is to ease as much as possible the access to the project's results for the general public and European research institutions.

However, our efforts for promoting the generation and management of open data in IOStack must continue and for this reason the current manuscript is not a definitive version of the IOStack's data management plan. In contrast, this document will evolve both *quantitatively* —number of open data items available— and *qualitatively* —enhancing the presented action lists, possibly including new actions— as the project progresses.

References

[1] E. Commission, "Guidelines on open access to scientific publications and research data in horizon 2020." http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf, 2013.

[2] Wikipedia, "Comparison of free and open-source software licenses." https://en.wikipedia.org/wiki/Comparison_of_free_and_open-source_software_licenses.

[3] Wikipedia, "Apache license." https://en.wikipedia.org/?title=Apache_License.