DRI Infrastructure Road Map 2018-2022
Version 1.8 For Publication
11 December 2018
Authors: Kathryn Cassidy, Lisa Griffith, Stuart Kenny, Kevin Long

Background to the Digital Repository of Ireland

The Digital Repository of Ireland is a trusted national infrastructure for Ireland's social and cultural data. We preserve, curate, and provide sustained access to a wealth of Ireland's humanities and social sciences data through a single online portal. The repository houses unique and important collections from a variety of national organisations including higher education institutions, cultural institutions, government agencies, and specialist archives.

DRI is a Trusted Digital Repository and was certified by CoreTrustSeal in 2018, having previously been awarded the Data Seal of Approval in 2015. DRI has staff members from a wide variety of backgrounds, including software engineers, designers, digital archivists and librarians, data curators, digital imaging experts, policy and requirements specialists, educators, programme and project managers, social scientists and humanities scholars. DRI was originally built by a research consortium of six academic partners working together to deliver the repository, policies, guidelines and training. Core academic institutions continue to manage the repository and implement its policies, guidelines and training. These are the Royal Irish Academy (RIA), Trinity College Dublin (TCD) and Maynooth University (MU).

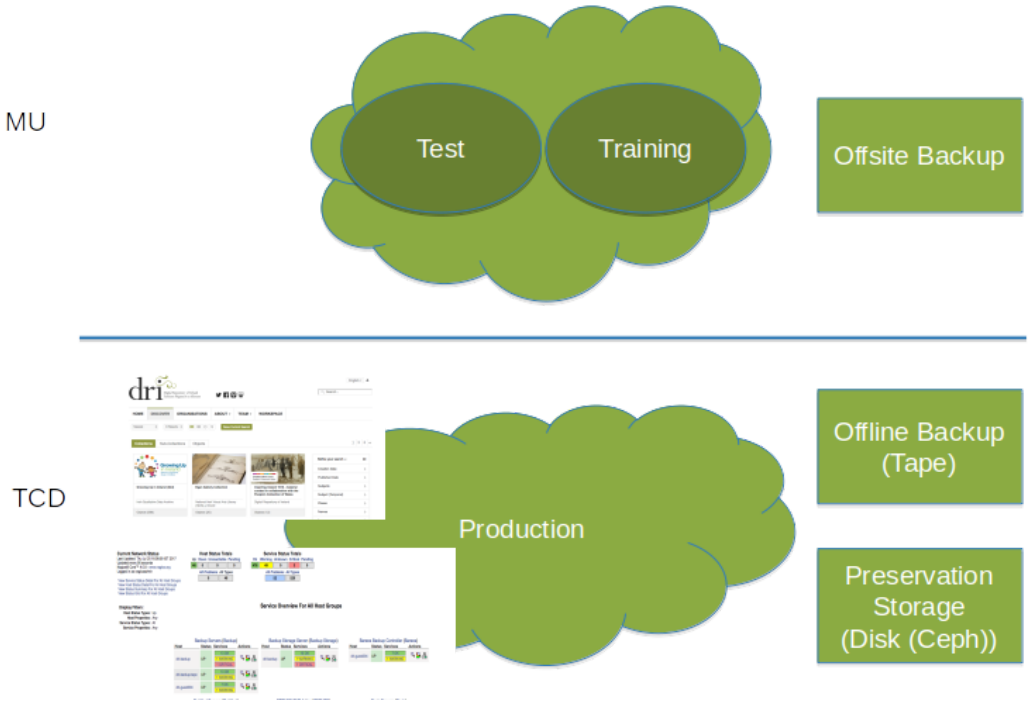
DRI receives a core operating grant from the Department of Education and Skills via the Higher Education Authority and the Irish Research Council, as well as funding from competitive research grants, projects, and philanthropic sources. DRI's membership structure ensures that depositors directly invest in the infrastructure's long term sustainability, while also helping to shape DRI's future via membership fora, direct feedback, and participation in DRI's governance structure.

Aim of the DRI Infrastructure Road Map: 2018-2022

The aim of DRI's infrastructure road map is to show how the repository is investing in the technical infrastructure to ensure the long-term stewardship of ingested collections. The road map also shows what funds are needed to sustain DRI's servers and related equipment. With the launch of DRI's membership model in February 2018 there are plans for a significant increase in DRI's ingested collections. This road map will help pinpoint the stages at which DRI needs to increase space and capacity. It also shows new and potential members how storage space will be created and supported to allow for the ingest of their collections from 2018-2022 and beyond. The road map also supports, if necessary at a future date, a re-evaluation of future space needs.

Summary of current infrastructure

The current infrastructure is divided between two sites: Trinity College Dublin (TCD) and Maynooth University (MU). TCD hosts the production services and offline backup storage, with training, test and offsite backup located at MU.



Both sites follow the same design in their deployment. A cluster of servers is used to provide the total compute and storage. The storage is delivered by Ceph (https://ceph.com), a distributed storage system. On top of the Ceph storage a private OpenNebula (https://opennebula.org) cloud is used to run the majority of the services as virtual machines.

At TCD, ten servers comprise the Ceph storage cluster, providing approximately 268TB of available storage. A further ten servers are used for the OpenNebula cloud services, providing the CPU and memory needed to run the virtual machine services.

At MU, six servers are used for both the storage and private cloud, with approximately 26TB of storage available.

TCD also houses the infrastructure used for backup. As per DRI's backup policy we keep three copies of the data stored. Data is first written to disk on a server located separately to the main Ceph storage cluster. A second copy is written periodically to offline tape storage. The third copy is stored offsite by utilising the storage space available at MU.

Further details on the infrastructure and how it is deployed is given in the DRI Infrastructure Report [INFRA-REPORT]. For more on how data is stored for preservation refer to the DRI Preservation Policy [PRES-POLICY].

Future collections and their needs

There are two factors that will significantly affect the DRI's housing of future collections: increased membership and new data types. We expect to grow our membership annually. This will result in future member institutions depositing data in addition to current members growing their collections. DRI also expects, and is encouraging, a greater intake of research data, which could include models, code or even executable programmes. We also anticipate an increase in specialist data sets including 3D models and spatial data such as GIS based projects and shapefiles. The exact nature of research data sets is difficult to predict in either volume or complexity, varying from tabular and text based data to the larger more complex spatial varieties specified above. The ingest of such data sets will also likely require modifications and development to DRI's front end in order to facilitate the best use of the data sets. This expected increased intake in data, including new formats, makes it difficult to estimate the precise storage requirements for the future, but amongst many smaller ingests, we predict the deposit of at least two very large collections a year in addition to our current holdings.

Additionally DRI expects to see increasing volumes of AV and audio content based on current enquiries from both media archives and oral history organisations. Such data can be vastly larger than text/image based collections and could exponentially affect storage requirements.

Future needs

It is necessary to plan for both obsolescence of current hardware and infrastructure, and expansion of the current capacity to ensure that storage and compute resources keep pace with demand as DRI's membership and collections grow.

A priority for 2018 is to replace the original backup system which has reached its storage capacity with newer servers. The current backup server in Trinity College should be replaced with a larger capacity machine, and a secondary backup server should be purchased for the Maynooth University site. Each of the backup servers is expected to cost c. €6,000 (VAT inclusive).

DRI operates a disk-to-disk-to-tape (D2D2T) backup strategy. Backups reside not only on disk on the backup servers, but are also copied to tape storage. The current tape backup system is out of warranty and should be replaced. In order to keep costs low, an alternative is to colocate DRI's tape backup with the Trinity College Research IT group's managed backup system. Research IT has agreed that this would be possible, although it would be necessary to purchase a number of tapes in order to increase capacity for this purpose. This would realise a significant cost saving over purchasing a dedicated tape backup system, without compromising DRI's ability to control its backup strategy.

DRI considers these upgrades to the backup system as the highest priority, as they are vital to reliable long term preservation of DRI's data.

A number of other machines are either out of warranty or will be out of warranty by end 2019. The out-of-warranty servers can continue to be used, but it is advisable to gradually replace these so that there is always a core level of capacity available to run a minimal site should those out of warranty servers fail. As more machines are replaced, some of the older out-of-warranty machines can be migrated to the secondary site at Maynooth University, or migrated to less critical functions within the main site (e.g. running support services such as build and test tools, rather than production services).

Four OpenNebula servers which run the various DRI Repository services in the primary site are currently out of warranty. It is proposed to gradually replace these on a phased basis spreading the cost over a number of years. The first of these is slated to be purchased in 2019.

In terms of the primary storage, six of the current storage servers are due to go out of warranty at the end of 2019. A plan has been put in place to gradually replace these in order to maintain existing capacity. Because of the falling cost of storage, current server offerings will likely provide higher storage capacity for the same price than was available when the existing servers were purchased in 2014. At least one new server should be purchased by end 2019 with a view to replacing the current storage capacity of two of the existing servers. After 2019, it is expected that we can gradually replace the rest of the servers at a rate of two new machines per year.

At present, the available storage is well above actual usage, and the phased replacement of storage servers from 2019 onward is likely to increase the overall storage capacity. Given our projected membership numbers and the predicted rates of ingest, this type of organic capacity growth is expected to be sufficient for the next two years. By 2021 we hope to be ingesting larger volumes of research data which will require a step change in storage capacity. Some additional investment may be required in 2021 to allow for the purchase of additional storage servers to grow the overall capacity, in addition to ongoing replacements for ageing and out of warranty machines. This storage growth requires maintenance and ingest support, so new human resources will also be required to grow at scale to support Ireland's Open Science participation and responsibilities.

Although several servers at the secondary site at Maynooth University are also out of warranty, it is considered less critical to replace these with new machines. Instead, as machines at the primary site are replaced, some will be moved to MU to replace end-of-life hardware, and also to gradually increase capacity in that site. This is an effective and efficient use of resources.

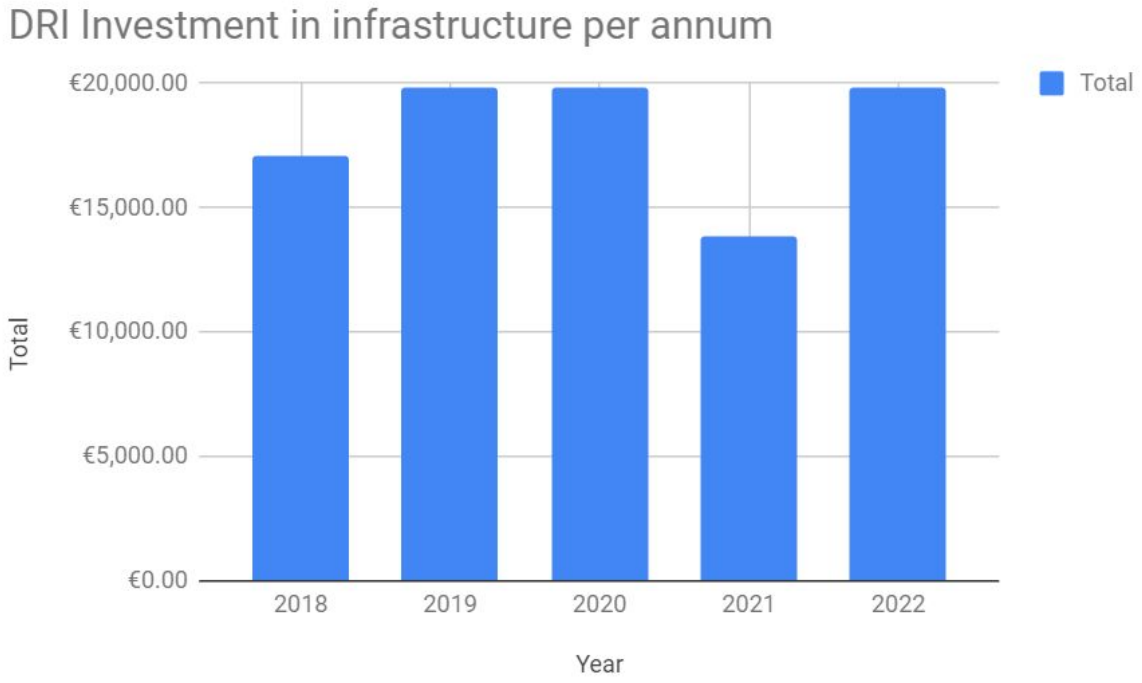
In addition to the servers, ongoing upgrades to the network infrastructure are required. Much of this work will, however, be undertaken and financed by the Trinity College Research IT group, and the DRI's use of their infrastructure will give us the benefit of this investment without having to commit additional resources.

Additional technology investment may be required from time to time to allow us to test and roll out new or experimental services and features. An annual review of these additional technical requirements will be made with funds ring-fenced to €2,000 per annum (with some scope to adjust this budget depending on needs). Where appropriate, this investment will be supplemented by securing project-based funding.

By the end of the 5 year period we will have replaced our existing backup system and the out-of-warranty OpenNebula and storage machines. We will also have a number of additional servers, expanding capacity and adding new functionality.

Investment

To meet the infrastructure needs outlined above, as well as to expand our capacity for the future needs of our members and in order to grow our investment DRI is planning an investment of €90,500 between 2018 and 2022. This investment will be phased and it is projected that it will take place in the following way:



This is an annual average investment of €18,000 based on current projections of membership growth. This cost will allow DRI to replace its current infrastructure as it ages and goes out of warranty. It will also allow DRI to expand its capacity and to meet the future needs of DRI's members. It also provides a contingency to meet future unknown costs.

Road Map 2018-2022 and beyond

This road map was developed in Autumn 2018 based on the needs of early members who signed up to DRI's membership model. As such, the detail here is based on forecasted needs of new and potential members. DRI is continually tracking the storage and capacity needs of collections and its members so this road map will be reviewed in 2020 to ensure that these projections continue to be fit for purpose.

References

[PRES-POLICY] Digital Repository of Ireland. DRI Preservation Policy, Digital of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.2r377c523

[INFRA-REPORT] Digital Repository of Ireland. Building the Digital Repository of Ireland Infrastructure, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.qr474f68n