IST AUSTRIA
Institute of Science and Technology

Document Name Implementing the institutional data repository IST DataRep

Author Barbara Petritsch
Version 1.0
Published 26.06.2017
Location https://repository.ist.ac.at/

Table of Contents

Open Access to Research Data @ IST Austria
Implementing the institutional data repository IST DataRep

Preparation, Implementation, Operation and Outlook
2012 – 2017
Project Report

Abstract

In this report the implementation of the institutional data repository IST DataRep at IST Austria will be covered. Starting with the research phase when requirements for a repository were established, the procedure of choosing a repository-software and its customization based on the results of user-testings will be discussed. Followed by reflections on the marketing strategies in regard of impact, and at the end sharing some experiences of one year operating IST DataRep.

Introduction

At IST Austria founded in 2009 research in the fields of life sciences, formal sciences and physical sciences is conducted. The young scientific institute is expected to grow up to 90 research groups in 2016. At the time of the repository project starting 19 research groups were based at IST Austria.

Anticipating the growing need of research data management (RDM) including Open Access to data regarding specificly to the changing requirements of funding agencies IST Austria decided to engage in implementing an institutional data repository in 2012 and in 2015 the IST Austria Library implemented IST DataRep.

Preparation

Survey
Actual State of Research @ IST Austria[1]

The reason for the survey was to get an idea of data creation and data handling at IST Austria. The questionnaire was compiled with the future repository in mind: What kind of data is created at IST Austria? How much storage space is needed for a certain amount of time? Are there any special requirements for the metadata? How do scientists at IST Austria store, share and archive their data? The survey was distributed online via Google Docs. Of all 19 research groups one designated

member was invited to participate. The small scale of the survey (n=19) allowed for us to design a questionnaire with free text response options.

Evaluation

The main finding of the survey was the importance of offering researchers a possibility to publish their research data directly in an institutional repository. At this point in time the individual strategies of the researchers did neither assure secure storage over a long time period nor did they support publication of data as requested by the funding agencies.

Another, expected result was that at IST Austria a high variety of data formats is created. Certain data are suitable for recognized subject repositories, which we recommend to use because they are most likely better equiped to meet the requirements. For the so-called long tail of research (data)[2] at this time an institutional data repository was the best choice.[3]

Institutional resources

At IST Austria half an FTE (full-time equivalent) was provided for the project, divided equally between IT and the library. No additional staff had been hired so this can be considered a rather rough assessment for the estimated project period of two years. In numbers this means 1600 working hours. There is only fragmentary documentation of working hours, which doesn't make the numbers particularly meaningful.

Approach

An institutional data repository's main tasks are to collect, disseminate and archive research data. Based on the data life cycle (Pic.1) we decided to start focussing on the dissemination aspect – giving access to data. Therefor data deposit is not manadatory but voluntary at the institute. The aspects of long-term storage and archiving were neglected at the time but focused on afterwards.

Pic. 1: Data Life Cycle adapted from the UK Data Archive

Don't reinvent the wheel

There are numerous initiatives and projects in the field of research data management (Pic.2). Those were sources of know-how but also projects we could get involved in. National as well as international and global operating projects are worth exploring. Sometimes one or the other emerges (or disappears) and others merge. So it is advicable to keep track with current developments.[4]

Pic. 2: Initiatives and projects in the field of RDM (updated 2016)

Repository Software

Pic. 3: List of considered repository applications

With the help of the gathered information a list of requirements for a repository software was drafted and ranked[5].

During the selection procedere we considered open source as well as proprietary software (Pic.3). From the beginning on the open source options were favoured out of little institutional resources but mainly because the proprietary options were designed strictly for publications. So these were the first to be ruled out (Content dm, Digital Commons, Digitool, Open Repository, Equella, Vital, Zentity[6]). One essential requirement was that the repository must support diverse data types and formats (i.e. Dataverse accepts various but only supports tabular and statistical data), Another claim was the support of standards (i.e. SWORD, OAI-PMH) to facilitate efficient and seamless workflows. Regarding the available ressources it was important for us to work with an application, which is easy to implemenet and to administrate (Dspace and Fedora are very

complex in their implementation). Last but not least it had to support the metadata schema of DataCite. At this point none of all the evaluated repositories fit all the categories but the University of Essex was piloting a plugin for Eprints called ReCollect,[7] which made Eprints the repsository software of our choice.
Other noteworthy characteritsics of Eprints are that there is a big user community due to the fact that it is one of the most common repository applications and it has a GUI (Graphical User Interface). Not to forget our IT department was already familiar with the application because the institutional publication repository (IST PubRep) implemented in 2013 is running with Eprints as well.

Implementation

User Testing and Adjustments

At first Eprints and the plugin ReCollect were installed and after that the first user testing was carried out. The specific setting was composed of a scientist and two librarians: one leading the dialog and the other taking minutes. The researchers went through the whole process of data deposit on their own. Du to the known usability of figshare the researchers were asked to also test its features and name their favourite ones. This turned out to be very helpful regarding customizing our repository. With the help of the protocolls a list of suggested adjustments was compiled, ranked, and were/are being implemented according to their priority and technical feasibility.
There have been 90 IT tasks and roughly 200 hours of development to adjust the application and fix bugs. After the first round of user testing further adjustments were made.

Metadata Fields

The users were asking for a tool, which is easy to use and enables a quick deposit. According to this request we designed an entry form, which consists of a considerable amount of prefilled fields (Table 1). Certain information is automatically selected and assigned i.e. account information, recommended presettings, or institution-specific generic information.

DataRep field	Preset
Creator	Name derived of account information
Contact address	Email address derived of account information
Resource language	En
Publisher	IST Austria
License	CC-0
Document type	Automatic recognition

Table 1: Presets in IST DataRep

In general all metadata fields were evaluated and those, which were rated negligible are blanked out, mandatory fields (e.g. in regards to Data Cite's Metadata Schema) were defined and some were even added (e.g. a second DOI field was added so there is one for the dataset itself and one for the related publication).[8] Some fields are optional even when the information is considered important but is not applicable to all data.[9]

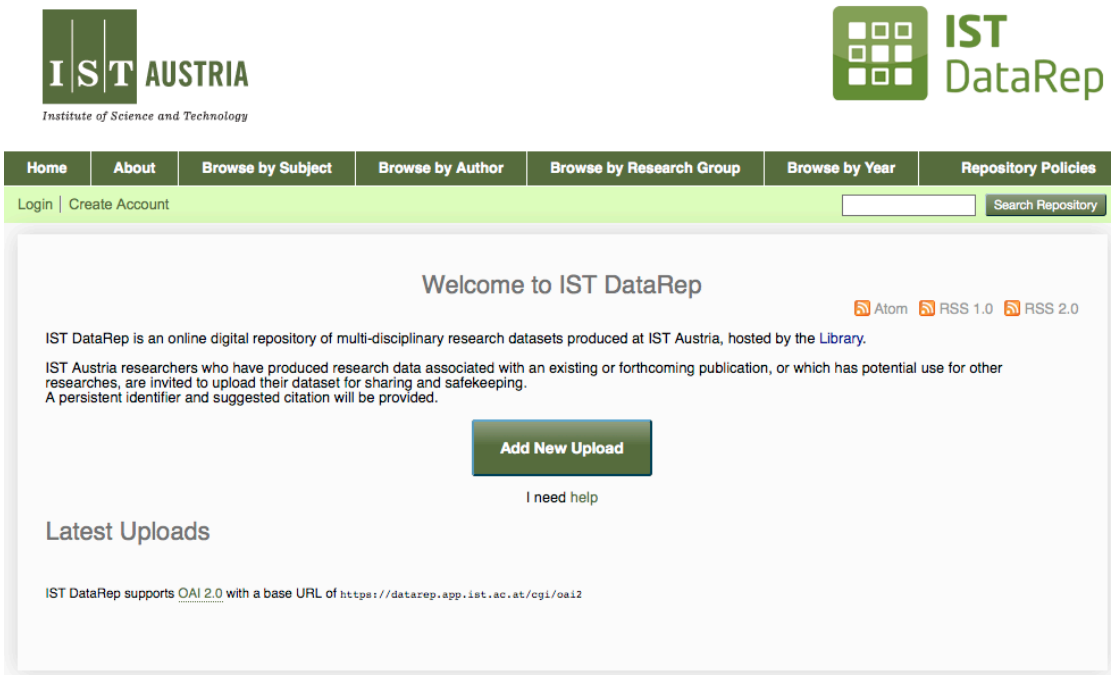
Field Help

In Eprints the field helps are hidden by default and are displayed by clicking the questionamark in the upper right corner. We decided to make them first of all visible and to alter the phrases in order to make them easier to understand and to address the users directly.

Pic. 4: Collection description field: default and asaption

User Interface

For inhanced usability a prominent ADD NEW UPLOAD button was placed on the repositories platform (Pic.5). From there the users are directed to their working area where they can upload, edit and deposit their dataset. This adjustment was drafted in anticipation of the single sign on system, which will be fully implemented in 2017. At the moment users still have to go through a separate login.



Pic. 5: Screenshot start page repository

Subjects

By Default Eprints is operating with the subject classes of the Library of Congress. We changed that to the Dewey Decimal Classification (DDC). There are several reasons why: even though we wouldn't get a DINI certification we still tried to allign with their requirements for the publication repository, one of them being to use DDC classification. After we used it for the publication repository slimmed down to those classes, which are represented at IST Austria we used the same set of classes for the data repository.

Operation

IST DataRep went live in August 2015 and is currently in its pilot phase. After one and a half years there are 17 datasets in the repository. This might sound measly but with the help of this deposits we made modifications to the application and its user interface to guarentee smooth processing during the upload. We also feel well prepared for more frequent uploads, which we expect after the amandments of the Horizon2020 Open Access policy for research data for 2017.[10]

Permanent Identifier

As already mentioned before we decided to go with DataCite a non-profit organisation providing DOIs especially for research data. At the moment we are minting them via DataCite's Metadata Store manually. There is an API in place for automatic registration but we didn't make it work until now. At the moment it is not a very urgent issue but it's planned for the IT development period of 2017.

EUDAT[11]

Right now we have a backup solution running provided by the institutional IT department. For an off-campus solution we are collaborating with the EC project EUDAT as one of their data pilots. The services we are interested in are B2Safe and B2Find, which are for long term archiving and enhanced visibility. Both services are implemented so far.

B2Safe

We are working together with the Karlsruher Institute of Technology (KIT) for long-term storage of our data collections. We agreed on the export of our data via XML packages, which Eprints offers as a feature. After the server was set up at KIT we tested to transfer XML packages from IST DataRep to KIT. To name some tasks the default script for the data export had to be adjusted to our requirements that each data collection gets its own package and not all data collections are pooled in one (big) package, A script for activating the export had to be created, to avoid duplicates it was important to make sure that metadata updates will simply replace the old package at the KIT servers. We also had issues with restoring the data collections from the XML packages, which was an Eprints bug fixed with features of a more recent Eprints version. Another bug was the failing of exports of files bigger than 100 MB due to time out.

B2Find

The implementation of this tool was a straightforward process. Via the OAI-PMH the metadata is collected and made available in the metasearch engine B2Find. We only had to adjust the protocol insofar that the DOIs are indexed and therefor represented in the metadata record.[12]

Being one of the data pilots it is not definite how or if this collaboration will continue. Aspects like costs, service agreement, etc. haven't been discussed yet. The development of the project itself will be of great influence on future collaborations.

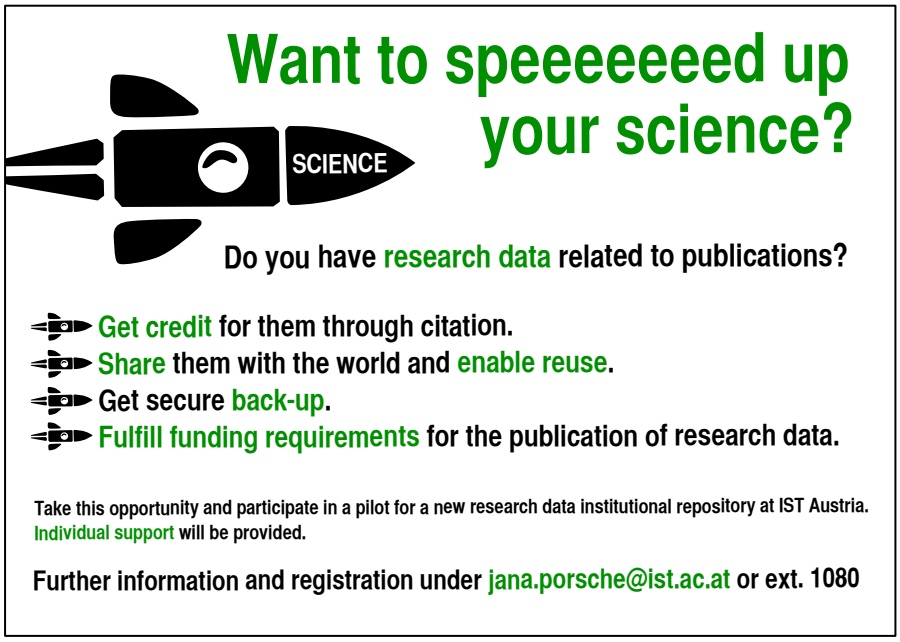
Research data management (RDM)

In 2016 we introduced RDM to our scientists focusing on outlining the general aspects of RDM. One-to-one meetings are offered for specific questions. RDM is also a field we can profit a lot from talking to our scientists face to face and learn about their habbits but also field specific traditions. This is especially thought as an exchange of knowledge both parties gain of. The objective is to adapt our services to the requirements in the research groups and support them with useful tools and information.

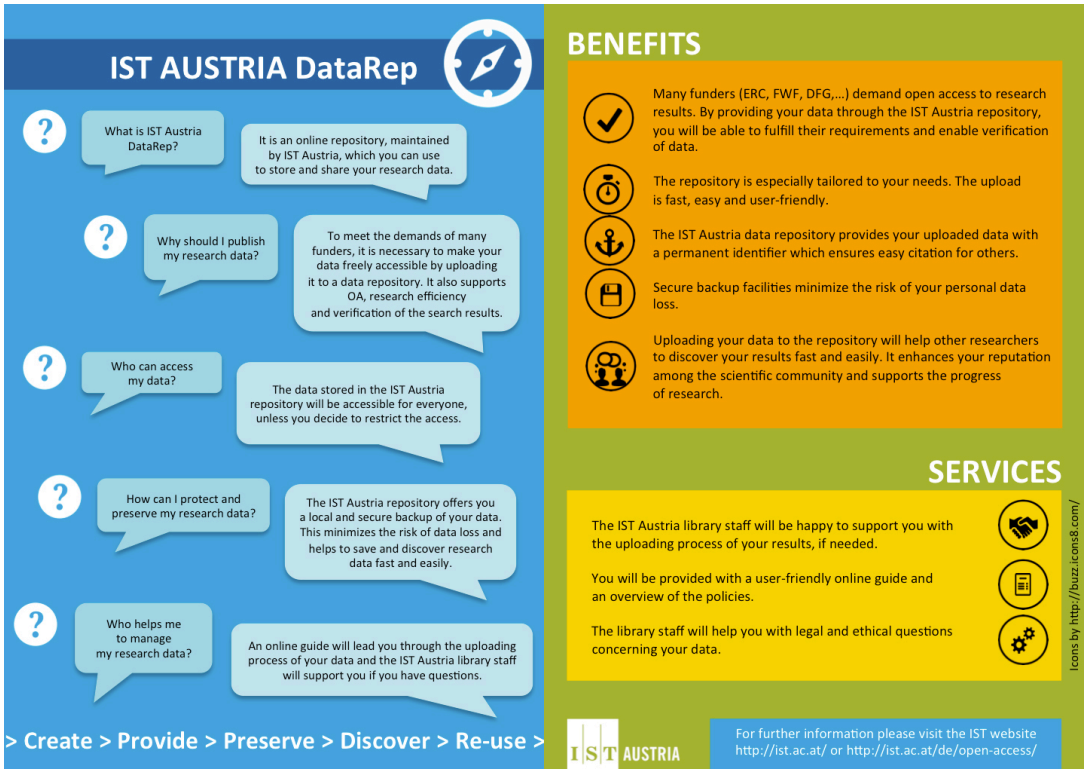
Marketing

In the course of branding the IST Austria library services a logo for IST DataRep was designed. In terms of awareness for data publication as a whole and for the IST DataRep in specifc posters were designed to find scientists at the institute willing to participate in the data repository pilot. Additionaly a project of the University Course of Library and Information Studies at the University of Vienna dealt with the task of raising awareness for data-publication and the data repository. A factsheet designed as a flyer was composed and an animation video was created for the same reason. The most effective strategy though was to approach institutional “champions“ in the field of open access publishing and get them on board. These people had already a close relationship with the library regarding this matters or were seeked out via the institutional publication database. At that time data publication was (perceived as) a minor aspect of scientific communication.

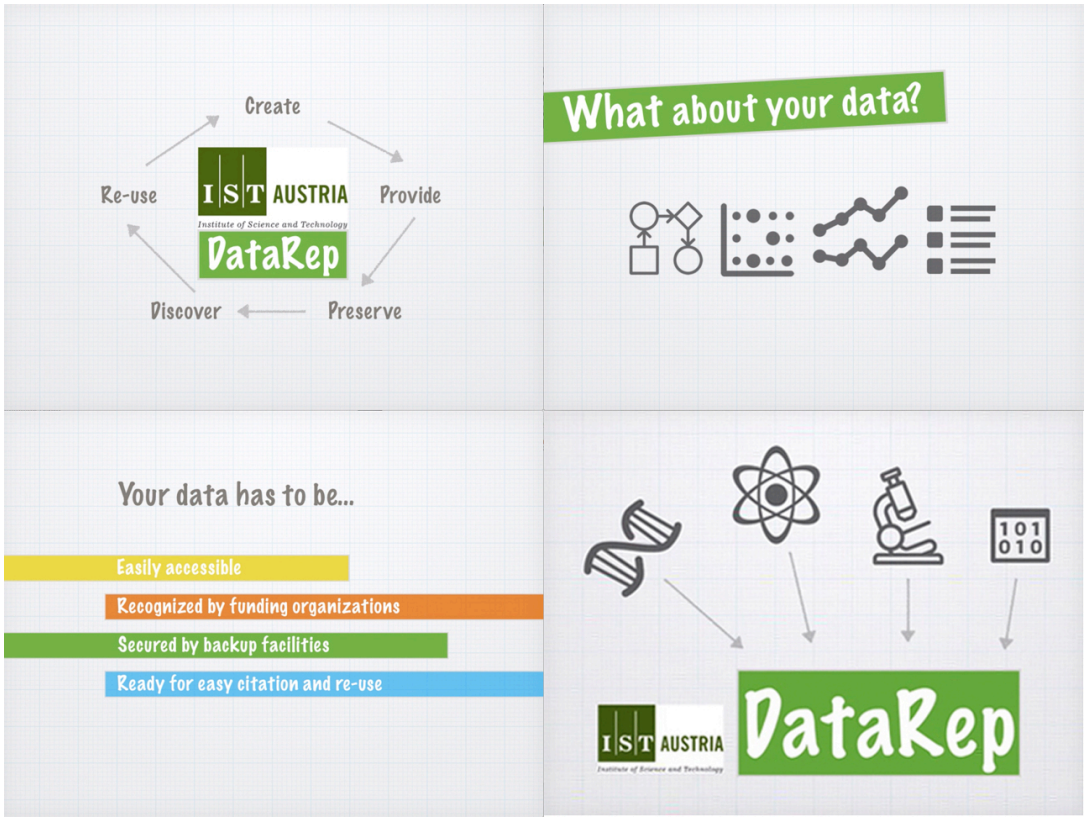
Since the European Comission announced the Open Data Pilot (2016), which states mandatory data publication this is changing, at least for the researchers at IST Austria because roughly half of the research groups receive funding by the European research Council.



Pic. 6: Poster for the data repository pilot



Pic. 7: Flyer IST DataRep



Pic. 8: Filmstills of clip promoting IST DataRep

Review and Outlook

Jump in at the deep end

Our approach was kind of contrary to the common practise and we implemented the tool before we set the rules or framework for operating it. For the library it was crucial to create a tool that is usable, understandable, practical and tailored for institutaional needs. One main idea to the project was to focus on only one challenge at a time. Our first priority was data publication. After the repository was available for data deposit/publication we started aiming at the next step: long-term storage. Right now we are finalizing a repository policy, which also will enable us to register with Re3data.org and consequently tackle OpenAIRE-compliance to report Open Access data publication for ERC grants. Instituionwide we are drafting a data policy, which will reference and relate to IST DataRep. Furthermore we are working on seamless workflows for mid-data (TB capacity) together with IT and two research groups who raised this issue.

In our opinion it is important to always be prepared to adjust to developments regarding either the requirements on the funders' or the researchers' side, technical innovation, changes in scientific communication, etc. and meet the challenge of bringing that all together. It will only be perfect for a second, it will only be beautiful for a few days but it will be changing all the time.

[1] For a detailed report see: Porsche, J. 2012. "Actual State of Research Data@ ISTAustria. Online: https://repository.ist.ac.at/103/.

[2] Tail (small, medium) data is in contrast to head (big) data described as heterogeneous, hand generated and created and processed under individual procedures. Therefore the curation isn't likely to be organised centrally and/or automatically (i.e. subject repository). Further more the accessibility is a major issue out of legal uncertainties and lack of effective platforms (i.e. institutional repositories). See also: Heidon, Bryan P: *Shedding Light on the Dark Data in the Long Tail of Science*. Library Trends 2008, 57, 2, 280-299. Online: http://hdl.handle.net/2142/10672

[3] At this point in time the general purpose data repository Zenodo (https://zenodo.org/) wasn't an option. Zenodo was launched in May 2013 by OpenAIRE and CERN.

[4] For a detailed list of initiatives in 2013 see: Porsche, Jana. 2013b. "Initiatives and Projects Related to RD." Online: https://repository.ist.ac.at/113/.

[5] The detailed report: Porsche, J. 2013. "Technical Requirements and Features." Find online: https://repository.ist.ac.at/135/.

[6] Zentity a Microsoft product was launched in 2009 but as of 2011 ceased to get any further development. An example for one of the disadvantages of commercial applications: If the company decides to discontinue development and the customer doesn't have access to the source code and no right to (re-)use it anyway.

[7] *The ReCollect plugin transforms an EPrints install into a research data repository with expanded metadata profile for describing research data (based on DataCite, INSPIRE and DDI standards) and a redesigned data catalogue for presenting complex collections. Developed by the UK Data Archive and the University of Essex, as part of the JISC MRD Research Data @Essex project.* See also: http://bazaar.eprints.org/367

[8] Recent developments showed that the DOI field for related publication should be repeatable because - and that's actually what the community tries to achieve - one dataset can be the base of several publications.

[9] For a detailed report see: Petritsch.B. 2017. "Metadata for research data in practise" in: ……. Info kommt noch

[10] Article 29.3 Open access to research data in: European Research Council. Mono-Beneficiary Model Grant Agreement. ERC Starting Grants, Consolidator Grants and Advanced Grants. Version 4.0; 27.2. 2017. Online zuletzt abgerufen am 16.5.2017 unter: http://ec.europa.eu/research/participants/data/ref/h2020/mga/erc/h2020-mga-erc-mono_en.pdf

[11] For further information on the scope of the project see: https://eudat.eu/what-eudat.
[12] See also: EUDAT services to guarantee long time archiving and visibility to the repository of IST Austria. Online via: https://eudat.eu/communities/eudat-services-to-guarantee-long-time-archiving-and-visibility-to-the-repository-of-ist.