

RD A GEDE Webworkshop

Adaptation of Repositories to the Digital Object Interface Protocol

22.5. 2019 from 16.00-18.00 CEST

URL: <https://global.gotomeeting.com/join/177773245>

Experts

Giridhar Manepalli (Director of Info. Mgmt. Technology, CNRI, Virginia)

Christophe Bianchi (Executive Director DONA Foundation, Geneva)

Rob Quick (Assoc. Dir. Science Gateway Research Center, Indiana University)

Paul Trilsbeek (DOBES Archive Manager, MPI for Psycholinguistics, Nijmegen)

Peter Wittenburg (MPCDF, Garching/Munich)

Questions

Why should well-established repositories switch to supporting DOIP?

What are the challenges for adapting to DOIP?

How much effort does it take to adapt to DOIP?

Which new possibilities will be opened?

The use of FAIR Digital Objects to implement the FAIR principles and to make data practices more efficient is now widely accepted. It is also widely agreed that repositories are the care takers of FAIR Digital Objects, i.e., they need to store, manage and curate DO's bit sequences, maintain globally resolvable PIDs for the DOs and manage metadata of different kinds describing the DO's bit sequence. The DO Interface Protocol is the unified protocol that allows to create and access FAIR DOs independent of how a repository is organising and modelling its data. Thus, repositories should talk DOIP, but reality is different.

- Repositories have been set up in many research/data infrastructures.
- These repositories have chosen specific technologies for storing and organising their data.
- These repositories have built more or less complex software to meet the major tasks such as allowing authorized users to ingest and access data.

The above-mentioned questions need to be answered urgently, since an adaptation will require some efforts and thus costs, since in some cases non-trivial challenges may occur and since repositories need to understand why they should take this effort. Therefore, we will conduct a first workshop via the web with key actors to first discuss the adaptation challenges and at the end discuss ideas about possible benefits.

To discuss the adaptation effort, we will take two examples of well-maintained repositories (in follow up workshops more examples could be discussed): The first example is being maintained at Indiana University and the second at the MPI for Psycholinguistics. In the Web-Workshop the experts will first explain the setup of the repositories and then discuss with the DOIP experts what best should be done to adapt the repositories to DOIP. Participants can also ask questions or give comments. The purpose of this meeting is to understand the challenges of adaptation and the degree of effort needed for such adaptation, for these two cases. In the case of IU, the adaptation work has already

started under the aegis of an NSF grant; in the DOBES case, there are no funds yet to do the adaptation which is probably the situation for most of the thousands of existing repositories.

The detailed discussion about the adaptation of these examples will help to estimate the efforts and costs for other cases. At the RDA plenary in Helsinki we intend to discuss other cases in detail.

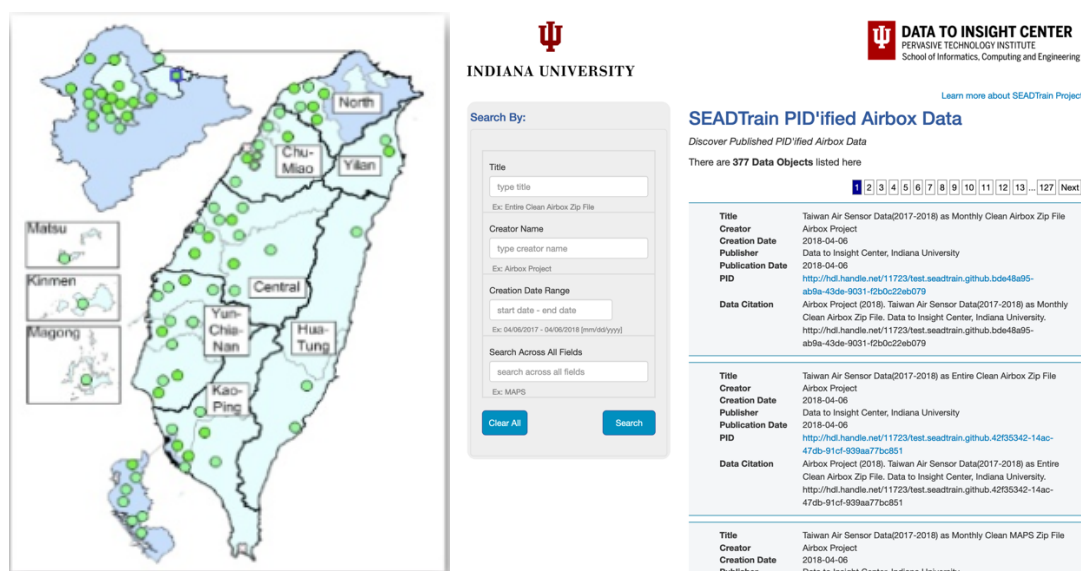
More documentation will be made available before the Web-Workshop at the GEDE-DO site.

Join this Working Meeting!

Sustainable Environmental Actionable Data Training (SEADTrain)

(<https://github.com/Data-to-Insight-Center/SEADTrain>)

Indiana University's SEADTrain project uses real-life environmental data collected from AirBox sensors deployed throughout Taiwan, Microsoft Azure storage, and the Robust Persistent Identification of Data (RPID) Testbed to provide a training platform for data science students.



The screenshot displays the SEADTrain web interface. On the left is a map of Taiwan with green dots indicating sensor locations across various regions like North, Central, and South. To the right of the map is a search interface with fields for Title, Creator Name, Creation Date Range, and Search Across All Fields. Below the search fields are 'Clear All' and 'Search' buttons. On the far right, under the 'DATA TO INSIGHT CENTER' logo, is a table titled 'SEADTrain PID'ified Airbox Data'. The table lists data objects with columns for Title, Creator, Creation Date, Publisher, Publication Date, PID, and Data Citation. The first entry is 'Taiwan Air Sensor Data(2017-2018) as Monthly Clean Airbox Zip File' with a PID of 'http://hdl.handle.net/11723/test.seadtrain.github.bde48a95-ab9a-43de-9031-f2b0c22eb079'. The second entry is 'Taiwan Air Sensor Data(2017-2018) as Entire Clean Airbox Zip File' with a PID of 'http://hdl.handle.net/11723/test.seadtrain.github.42f35342-14ac-47db-91cf-939aa77bc851'. The third entry is 'Taiwan Air Sensor Data(2017-2018) as Monthly Clean MAPS Zip File' with a PID of 'http://hdl.handle.net/11723/test.seadtrain.github.42f35342-14ac-47db-91cf-939aa77bc851'.

PIDs are assigned to SEAD data upon collection by the RPID testbed and the user interface (pictured above) allows high-level sorting and retrieval of data. While this sorting is important to reduce data wrangling tasks, the long-term goal is to use the Digital Object Interface Protocol (DOIP) to allow direct operational interaction with the data objects stored in the SEADTrain repository.

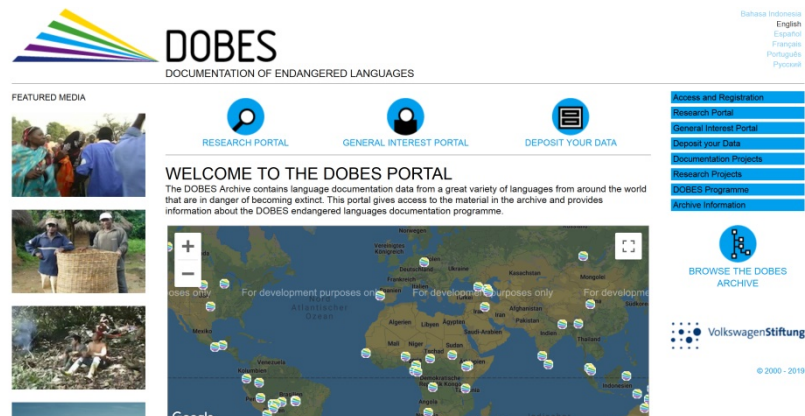
Development of a mapping service that will allow the SEADTrain data to be used without repository refactoring is underway. This mapping will integrate the existing data objects and kernel information resolvable from RPID with the goal of allowing DOIP defined operations. Upon completion of this research project SEAD will evaluate using the PID-based interface for additional data analysis needs as well as providing an easy interface for addition, deletion, and modification to the data repository.

DOBES Archive at MPI for Psycholinguistics

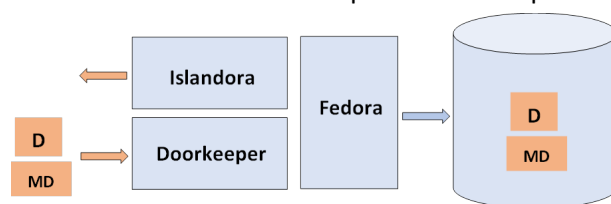
(<http://dobes.mpi.nl/>)

The DOBES archive has been set up from 2000 on and manages data of different types about roughly 90 endangered languages documented by 75 international teams that worked around the globe. Interested people can browse/search in the archive although access to some data such as about religious ceremonies etc. is restricted.

The DOBES Archive setup has been rebuilt recently and has been ported to Fedora Commons, a software library with repository functionality. From the beginning the DOBES technology team focussed on methods that would now be called "FAIR compliant", i.e., all data objects are assigned a PID (Handle) and metadata descriptions which are based on a community defined schema with registered semantic categories. The archive is built in form of a hierarchy of sub-collections to support easy management and curation, each of these sub-collections is basically FAIR compliant as well. Since the archive incorporates non-repeatable recordings of human heritage from each digital object 4 external copies to large computer centres are being generated and 10 remote archives are being supported. The metadata records are being exposed to service providers via an OAI-PMH port



Accessing the data or metadata is straightforward, since only the URL or PID is needed. For uploading new data and metadata a special doorkeeper module has been developed that does all the type



checking, some curation, requesting PIDs and generating the PID record, calculating checksums, updating the metadata, maintaining all the links, etc. An adaptation therefore would require an embedding or rewriting of the doorkeeper module. During the workshop we

want to work out how this could be done, what the possible challenges are and how much time it would cost.