
**ICSTI 2017 ANNUAL MEMBERS' MEETING & WORKSHOPS**
**4th April 2017 - Barcelona, Spain**

Barceló Sants Hotel
Plaça dels Països Catalans, s/n | 08014 Barcelona

# TACC Workshop

## 'Scientific Data Repositories : Use Cases, Innovation, and Best Practices'

ICSTI's **Technical Activities Coordinating Committee (TACC)** typically focuses on exploring and communicating technical aspects of innovative trends in information science-based tools that help make STI more useable and accessible. The **Chair of TACC is Brian Hitson**, Director US DOE/Office of Scientific and Technical Information (DOE/OSTI).

### Workshop Description

Research funding in recent years often comes with the condition to make some of the resulting data openly available. Scientific data can be made available through a number of mechanisms, including being hosted in general scientific data repositories. This workshop will explore various data repositories and data discovery tools, their plans for long term preservation and access to the data they host, provenance and updating/versioning of data, and cross-repository integration and search.

### Workshop Program

**Dan Valen** – figshare

***Active Research Data Management Workflows through to Preservation***

Academia has seen a dramatic increase in the number of data generated throughout the last few years and a similar jump in the number data repositories created to expose and store that content. With this growing awareness around data creation and reuse, in some cases urged forward by government policies and mandates around data sharing, we at figshare needed to provide certain assurances to our user base and to our partners in the academic community around data availability. As is often said of most infrastructure, the best infrastructure is invisible to the end user, and so we worked with service providers and our institutional partners to ensure that data published on figshare is

available for reuse by future generations of researchers via numerous preservation options. This talk will highlight different ways figshare and the archiving and preservation community are leveraging existing technologies to preserve research outputs on the platform.

***Dan Valen*** *joined figshare in early 2014 as its first US-based employee. As a product specialist, he focuses on the development of figshare in North America through community engagement, marketing and promotion, strategic partnerships, and educational outreach. Dan helps provide a lateral perspective across the research data management landscape in assessing the needs of researchers and institutions alike, while also offering guidance on current industry trends. In his previous roles, Dan spent over 6 years at one of the largest open access STEM publishers holding positions in editorial, trade publishing, and electronic content licensing.*

**Angelina Kraft** – Technische Informationsbibliothek (TIB), German National Library of Science and Technology

### *Establishing a generic Research Data Repository: RADAR*

While the research data environment has become heterogeneous and the data dynamic, funding agencies and policy makers push towards findable, accessible, interoperable and reuseable (= FAIR) research data. A popular issue of the management of data originating from (collaborating) research infrastructures is their dynamic nature in terms of growth, access rights and quality. The presented RADAR - Research Data Repository - service strives to make a decisive contribution in the field of long tail research data: On one hand it enables clients to upload, edit, structure and describe (collaborative) data in an organizational workspace. In such a workspace, administrators and curators can manage access and editorial rights before the data enters the preservation and optional publication level. Data consumers on the other hand may search, access, download and get usage statistics on the data via the RADAR portal. For data consumers, findability of research data is of utmost importance. Therefore the metadata of published datasets can be harvested via a local RADAR API or the DataCite Metadata Store.
Being the proverbial "transmission belt" between data producers and data consumers, RADAR specifically targets researchers, scientific institutions, libraries and publishers. RADAR possesses a modular software architecture based on the e-research infrastructure eSciDoc Next Generation. The data storage is managed by a repository software consisting of two parts: A back end addresses general tasks such as storage access and bitstream preservation, whereas the front end implements RADAR-specific workflows. Front end workflows include various data services: Metadata management, access control, data ingest processes, as well as the licensing for reuse and publishing of research data with DOI. Archival Information Packages (AIP) and Dissemination Information Packages (DIP) are provided in a BagIt-structure in ZIP container format. The RADAR API enables users to integrate the archival backend into their own systems and processes.

***Angelina Kraft*** *is team leader for research data management and scientific software associate at the German National Library of Science and Technology. She has a PhD in Biological Oceanography, which helps her work on assisting researchers better manage and share their data. Working in the Department of R&D, she co-manages the RADAR-Research Data Repository project, which aims to establish a generic data infrastructure. As a member of the DOI team she develops quality standards and best practices for the electronic publishing of research data.*

**Tim Smith** – CERN

## *Zenodo: A Research Data Repository for All*

Zenodo is open to research objects from any field. It was launched by OpenAIRE and is housed in the CERN data centre where it is founded on the storage infrastructure matured for big data. This talk will describe how Zenodo captures and serves research objects and the associated information necessary to continue making them accessible into the future. It will also describe preservation actions to guarantee the data fixity and experience of the needs of making it mutable, through update and versioning actions!

***Dr. Tim Smith*** *leads the CERN group that develops, installs and maintains instances of Invenio, the CERN Open Source Digital Repository system. He is heavily involved in initiatives to drive digital archives at the institutional and subject level and to populate them with content of a broad range of media types. He drove the launch of Zenodo within the OpenAIRE project as an open data service for the long-tail of science. He also drove the launch of the CERN Open Data Portal to share the LHC data with the world. Prior to these tasks he led teams responsible for computing farm management and physics data management. He holds a PhD in Physics and performed research at the CERN LEP accelerators for 10 years.*

**Amir Aryani** – Australian National Data Service (ANDS)

## *Research Data Switchboard*

Driven by the rapid development of data storage technology, the number of data repositories is growing fast. Researchers now have access to a range of data infrastructures such as discipline-specific repositories and national (regional) data infrastructures. The problem is that these infrastructures are often operating in silos; that is, they do not connect their datasets to related research information in other platforms.

One solution to this problem is the work undertaken by the Data Description Registry Interoperability WG of Research Data Alliance (RDA). The group has developed the Research Data Switchboard (rd-switchboard.org) which connects datasets and related information across research data repositories using information on co-authorship and jointly funded projects. The Switchboard enables traversing the graph of scholarly works from datasets to publications, grants, researchers and other datasets with multiple degrees of separation.

In this talk, we present an overview of the Research Data Switchboard and discuss how it can be used to link publications and datasets to the Research Graph -- a distributed graph of scholarly works derived by RD-Switchboard. Also, we will show a live demo of traversing the graph of connections between publications, datasets, researchers and research projects across Australia and Europe.

***Amir Aryani*** *is working in the capacity of a project manager for Australian National Data Service (ANDS), and he is the co-chair of the Data Description Registry Interoperability WG in Research Data Alliance. He has a PhD in computer science and the research background in software engineering and evolving software systems. He is the technical lead for the RD-Switchboard project, and he has a significant role in developing the open collaborative project called Research Graph that connects research data, publications, and other scholarly works to research projects and grants across global data infrastructures.*

**Ingrid Dillo** and 

**Peter Doorn –** DANS

## *FAIR Data in Trustworthy Data Repositories: Everybody wants to play FAIR, but how do we put the principles into practice?*

Research funding in recent years often comes with the condition to make some of the resulting data openly available. More and more data rapidly become available. Therefore, there also is a growing demand for quality criteria for research datasets.
In our presentation we will position DANS and its services in this context, with a special focus on our work in the areas of certification of data repositories and the operationalization of the FAIR principles.
We will argue that the DSA (Data Seal of Approval for data repositories) and FAIR principles get as close as possible to giving quality criteria for research data. They do not do this by trying to make value judgements about the content of datasets, but rather by qualifying the fitness for data reuse in an impartial and measurable way. By bringing the ideas of the DSA and FAIR together, we will be able to offer an operationalization that can be implemented in any certified Trustworthy Digital Repository.
In 2014 the FAIR Guiding Principles (Findable, Accessible, Interoperable and Reusable) were formulated. The well-chosen FAIR acronym is highly attractive: it is one of these ideas that almost automatically get stuck in your mind once you have heard it. In a relatively short term, the FAIR data principles have been adopted by many stakeholder groups, including research funders.
The FAIR principles are remarkably similar to the underlying principles of DSA (2005): the data can be found on the Internet, are accessible (clear rights and licenses), in a usable format, reliable and are identified in a unique and persistent way so that they can be referred to. Essentially, the DSA presents quality criteria for digital repositories, whereas the FAIR principles target individual datasets.

***Dr Ingrid Dillo*** *holds a PhD in history and has worked in the field of policy development for the last 25 years, including as senior policy advisor at the Dutch Ministry of Education, Culture and Science. Ingrid is now deputy director at DANS. Among her areas of expertise are research data management and the certification of digital repositories. She is a member of the Board of the Data Seal of Approval (DSA), the Technical Advisory Board of Research Data Alliance (RDA), and the Board of Directors of the DRYAD repository. She also is vice chair of the Scientific Committee of the ICSU/World Data System (WDS), co-chairs the RDA/WDS Interest Group on Certification of Digital Repositories and the RDA/WDS Interest Group on Cost Recovery for Data Centres, was an an active member of the former RDA/WDS Repository Audit and Certification DSA-WDS Partnership, and participates in the Research Data Expert Group of the Knowledge Exchange.*

Further information: https://pure.knaw.nl/portal/en/persons/ingrid-dillo(ad57b846-7aeb-4d86-9bf5-859973d7e00a).html

***Dr Peter Doorn*** *is director of DANS. He was co-founder of the Netherlands Historical Data Archive in 1989 and has been active in the domain of digital research data ever since. He is chair of the Science Europe Working Group on Research Data, national representative and vice-chair of the CESSDA General Assembly, former national representative of DARIAH ERIC, former chair of Research Data Netherlands, board member of the Research Data Alliance Organisational Advisory Board, and (board) member of various other national and international data-related organizations. He is editor of the recently founded Research Data Journal for the Humanities and Social Sciences.*

Further information: https://pure.knaw.nl/portal/en/persons/peter-doorn(9214e6f3-386c-4a3e-abfc-0535f2f52b5a).html

**William Michener** – DataONE

## *DataONE: Supporting Data Discovery, Sharing and Reproducibility to Enable New Science*

DataONE supports discovery, access, preservation, and sharing of data about life on Earth and the environment that sustains it. Scientists and decision-makers use a consistent and easy-to-use search tool for finding and accessing data that are stored in data repositories worldwide—the DataONE Federation. Researchers now have access to hundreds of thousands of data sets stored in dozens of major data repositories. The data represent observations from across the world's continents and oceans and are valuable resources for science and society. DataONE empowers scientists and others to easily discover and access data about life on Earth and the environment through a set of data management tools and services that are tailored to both researchers and data repositories. DataONE products encompass four areas that are targeted to meet individual and repository needs for: (1) Search and Discovery, including semantic annotation; (2) Research Data Management, including new provenance-tracking capabilities via enhanced MATLAB, R and workflow tools provided by DataONE; (3) Education and Training, including an online webinar series; and (4) Federated Repository Services, including data replication and usage tracking.

***William Michener*** *is Professor and Director of e-Science Initiatives at the University of New Mexico's College of University Libraries & Learning Sciences. He serves as Project Director for New Mexico's National Science Foundation (NSF) and Department of Energy EPSCoR (Experimental Program to Stimulate Competitive Research) Programs, and Data Observation Network for Earth (DataONE)—a large DataNet project supported by NSF. He is involved in research related to creating information technologies supporting data-intensive science, development of federated data systems, and community engagement and education. He has a PhD in Biological Oceanography from the University of South Carolina and has published extensively in marine science, as well as the ecological and information sciences. He serves on several Boards of non-profit organizations and has expertise in project management and meeting facilitation. He is Editor of Ecological Archives, Associate Editor for Ecological Informatics, and a member of the Editorial Board for Ecology.*

Contact Information:
College of University Libraries & Learning Sciences, MSC04 2815, 1312 Basehart Drive SE, University of New Mexico, Albuquerque, NM 87131-0001 USA
email: william.michener@gmail.com