C2CAMP Partners

Corporation for National Research Initiatives

MPCDF Max Planck Computing & Data Facility

CSC – IT Center for Science

Deep Carbon Observatory

ICOS Integrated Carbon Observation System

Indiana University Bloomington

CLARIN Common Language Resources and Technology Infrastructure

CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre

NCAR National Center for Atmospheric Research

DISSCO Distributed System of Scientific Collections

JÜLICH Forschungszentrum

Computer Network Information Center Chinese Academy of Sciences

doi

Natural History Museum

DIRISA

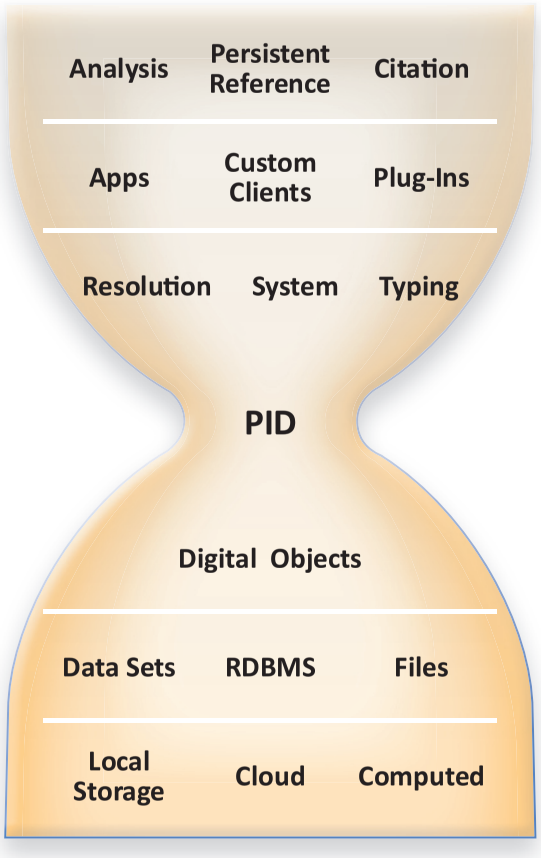
DKRZ Deutsches Klimarechenzentrum

is-enes Infrastructure for the European Network for Earth System Modelling

Interlinking Digital Repositories

Enabling efficient Digital Object Management

Digital Objects as Foundational Entities in the Global Data World



C2CAMP is an initiative that wants to implement a flexible and extendable testbed to test and integrate promising components for a global data infrastructure. It will widely rely on specifications worked out in RDA and other global initiatives (OAI, W3C, etc.). C2CAMP will be guided by the FAIR principles, improve the linking between repositories and create a world where greatest value can be extracted from Digital Objects to the benefit of science. In doing so, it will contribute to regional initiatives such as the European Open Science Cloud.

- The systematic use of stable PIDs will revolutionise digital object management practice and the supporting service architectures towards reliable referencing and establish the mechanisms for provenance tracking.
- PIDs will enable common access pathways to object metadata for both human users and automated processes at an ultimate level of interoperability. Concerns of data location, encodings and storage mechanisms can be separated from the actions that are of real concern to users, and dealt with much more efficiently and effectively.
- Agreeing on the principal architectural model to organise data, components and services at the virtualisation layer will allow us to implement automatic type-based workflows that will create the manageable data domain which we urgently need to cope with the masses and heterogeneity of data.

C2CAMP Team

Ari Asmi, Jonathan Clark, Paul Jessop, Dimitris Koureas, Werner Kutsch, Leif Laaksonen, Larry Lannom, Michael Lautenschlager, Thomas Lippert, Daniel Mallmann, Colin J. McMurtrie, Eric Nienhouse, Per Öster, Mark Parssons, Robert Quick, Raphael Ritz, Thomas Schulthess, Koenraad de Smedt, George Strawn, Dieter van Uytvanck, Anwar Vahed, Tobias Weigel, Peter Wittenburg, Tian Ye

Digital Object based Landscape in 10 years

With C2CAMP we suggest a global testbed project to demonstrate the great potential of the principles described above, to fit well into large programs such as European Open Science Cloud and to implement the FAIR principles (Findability, Accessibility, Interoperability, and Reusability). We will be able to interlink the various repositories all implementing different data models by offering a unique and generic architectural framework in which a common digital object management interface provides a comprehensive interoperability solution.

Within 10 years we expect a scenario that can be described by the following assertions:

Users and machines can operate in a global FAIR compliant domain of Digital Objects described by PID and metadata information. Users do not have to care anymore about the way data is organised and stored.

- The inherent virtualisation at a global level allows us to define and build the required registries and adapters to repositories.
  The network of data centres will implement an infrastructure that will support the high level actions efficiently, based on unified protocols.

FAIR
principles

Findability · Accessibility
Interoperability · Reusability

Digital Objects as Foundational Entities in the Global Data World

In general in computer networks, apparently meaningless self-routing messages between Internet devices are exchanged and then aggregated at the destination to serve a purpose.

Amongst the very first applications were email and file exchange.
A number of years later HTTP and HTML were introduced to build the WorldWideWeb which was characterised by its founders as a large scale hypermedia initiative to enable common access to a large collection of documents. In particular the introduction of the web required the notion of Uniform Resource Identifiers to be added on top of the IP numbering system. In practice, identifiers in the web identify an internet device and a locally interpreted path to a web-page or service.

It is the simplicity of the approach that provided enormous value and made the internet and the web a gigantic success. But we have had to learn to live with a number of limitations.

The most severe ones are the binding of identification with specific protocols and locations and the well-known instability of path specifications known as link-rot.

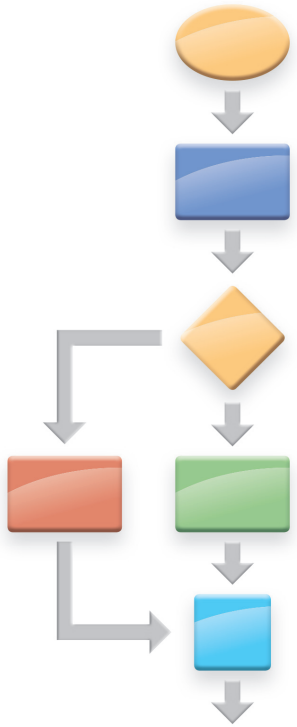
It is now common to try to manage and analyse data through web-services and interfaces. With the rapidly growing and vast amounts of geographically distributed and heterogeneous collections of databases and other structures of data, this approach is not sustainable, not scalable and results in large missed opportunity costs.

We believe that a new approach is required to solve the large set of challenges ahead, including:

- repositories may be based on any number of file systems, databases and clouds, and have highly different types of internal organisations,
- data will be moved and copied by automated processes to a variety of places, i.e., have different instantiations,
- there is an increasing need to trace data usage across life-cycle phases and to maintain provenance information,
- numerical data is not self-evident to either humans or machines and is opaque to those who did not create it without some agreed upon system of typing,
- an increasing number of references will be required in workflows, software systems, data structures and e-publications and they will need to be stable to guarantee (for example) reproducibility.

Digital Objects enable Automated Processing

In addition, automated processing of data of different types will be increasingly important to manage the streams of data and to extract useful knowledge that can be interpreted by humans. The domain of registered DOs opens the gate towards automatic processing.



DOs offer a unified interface for machines to carry out actions in a self-sustained way and to attain all kinds of metadata. PIDs are the keys to enable these paths to DOs independent from location or implementation system. In particular DO types associated with operations stored in “type registries” are powerful mechanisms to promote automated processing, not unlike the way in which MIME types today enable computer operating systems to automatically associate the correct applications with certain kinds of files. By embedding the most basic provenance recording tasks within automated processing and by promoting this to a core principle of the multi-layered approach, we can overcome the current barriers to gathering provenance information, which today often rely on user input that consumes their precious time and has proven hard to incentivise.

Typed Digital Objects combined with software agents that scan the data offered by repositories based on profiles can help ameliorate the data crisis.

Digital Objects enable Global Virtualisation

Millions of repositories have been setup during the last few years holding huge amounts of valuable data and all use different organisations of data and storage systems, be they files, clouds, SQL databases or no-SQL databases. We cannot expect that all these repository systems will be changed in order to introduce new approaches. Introducing, however, DOs referenced by persistent identifiers and resolved to meaningful metadata will allow us to define a common interlinking language and implement a global virtualisation. Developing adapters will make the various repositories part of a global domain. A PID can be resolved into directory paths, into hash values for a cloud or any kind of query for databases. In such a virtualised domain of DOs users should not have to care about the details of data organisation and storage, but have the ability to just interact with services that will be empowered by PIDs, well-structured metadata, types, collections which are themselves digital objects, all normalized to a uniform object layer.

Such an approach does not solve all the problems of data management and analysis, but if successful does reduce the current chaos of management systems that makes those problems much more difficult than needed and opens new ways to the difficult problems of knowledge creation and management. The next logical step to build this virtualisation domain is to develop the interface protocols that allow a) repositories to expose the DO they store by offering the corresponding PIDs (Resource-Sync, for example, would be a good candidate) and b) applications to access the bit sequences and/or the metadata belonging to a DO.

Digital Objects identified by globally resolvable PIDs form an interoperable virtualisation layer across all types of organisations of data. They can interlink storage systems of different kinds. The binding role of PIDs allows accessing bit sequences, metadata and other types of information which are crucial for re-usage.

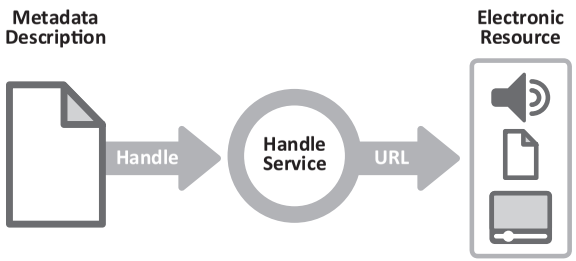
The concept of Digital Object

In 1995 R. Kahn and R. Wilensky introduced the concept of “digital objects” to overcome the limitations and work towards a manageable domain of data. This notion was taken up by well-known repository software builders such as Fedora Commons and D-SPACE and was one of the key pillars for clouds being introduced as object stores. In all these systems “digital objects” have a persistent identifier independent of any protocol or location and different types of metadata are associated with each object.

In current cloud systems it is a hash value that uniquely identifies an entity within the cloud and that is internally resolved to locations and metadata.

From about 2000 on, many initiatives world-wide who are operating with much data adopted the notion of DOs and developed architectures for managing Digital Objects using a global system for persistent identification.

After years of successful experience we can state that the concept of Digital Objects has shown its worth. Digital Objects have bit-sequences representing some content, are identified by globally unique persistent identifiers (PIDs) and are associated with different types of metadata. The PIDs can be used to refer to different types of information such as locations, checksums, types and other metadata to enable immediate operations.



Application
Layer

Users run applications such as collection builders, workflow engines, Virtual Research Environments, etc.

DO Representation
Layer

Users interact with logical representations of DOs and collections through interoperable operations, object properties and descriptive metadata.

DO Services
Layer

A layer of services offered by repositories and registries which curate the underlying logical descriptions and facilitate object actions

Basic Infrastructure
Layer

A network of federated computing & data centres store and process the DOs using a variety of different systems