High-Availability, Complementary Infrastructures for Persistent & Unique Identifiers for Data Objects & Published Collections based on Handle System

March 2012
The Data Creation Cycle

Data objects are being created in continuously larger numbers as the result of increasingly powerful sensors, as result of worldwide crowdsourcing activities and as result of computer simulations. They are being enriched as result of scientific activities - be it by manual intervention or by computer algorithms. Almost all of these data objects have a life-cycle, i.e. they are being created, validated, used, re-used, modified, moved and copied for various reasons, gathered into complex collections, etc. Data objects are annotated by content information, documentation and provenance information.

For many reasons it is so important to keep track of these data objects together with their annotations, to verify their integrity and authenticity and to quickly see the context of creation and the provenance.

With associating identities with each data object and each collection early at the beginning of their lifecycle and certainly at the moment where scientific publication refer to data collections we improve data life cycle management and access to data for true scientific purposes.

Handle System

Why is Identifier Resolution Important?

Resolution systems enable client software to go from an identifier to current state information about the identified object, such as where and how to access the object. Such identifiers can persist over changes to the identified object, such as changes in its location(s), ownership, and other attributes, persistence that is vital for maintaining data integrity over time.

What is the Handle System?

The Handle System is a general-purpose identifier resolution system that has been in place for many years. Identifiers in the Handle System are made up of a prefix and a suffix separated by a slash, e.g., 10.1594/PANGAEA.667386 or 4263537/5030. The prefix is used by client software to find the specific servers within the widely distributed resolution system that will be able to resolve the identifier. Existing identifiers can be easily structured as handles in order to take advantage of the resolution system.

Who Manages the Handle System?

CNRI designed, implemented, and currently administers the root level of the system, but the bulk of the resolution services are managed by the thousands of organizations, communities, government agencies, and businesses around the world currently using the system on a daily basis. This includes many academic and national libraries, scholarly journal publishers, scientific institutes, and other information management groups. CNRI is working with other major user groups, including EPIC and DataCite to create DONA (Digital Object Numbering Authority) to manage the Handle System into the future. The new organization will be governed by the DONA Board, which will include experts and stakeholders from around the world.
DataCite

**Mission**
DataCite is a non-profit organization formed in London on 1 December 2009. Its aim is to establish easier access to research data on the Internet, to increase acceptance of research data as legitimate, citable contributions to the scholarly record and to support data archiving that will permit results to be verified and re-purposed for future study. Just as research is global, DataCite is global, with member institutions offering services and advice directly where they are needed by the researchers.

**Characteristics**
DataCite brings together the datasets community to collaboratively address the challenges of making final research data visible and accessible. DataCite works on global standards for data citation and referencing of data from journal articles. DataCite works primarily with organizations that host data, such as data centers and libraries. DataCite runs a central open metadata repository of scientific record (http://search.datacite.org).

Citable datasets become legitimate contributions to scholarly communication and reward data sharing. The registration of DOI names for data sets allows direct and traceable citations of the data sets as individual scientific records.

**Governance & Costs**
Most of the DataCite services are free for end-users. Additional costs for the Registration of DOI-names depend on the individual member’s business models. DataCite is an official DOI-registration agency and has 16 members from 11 countries. European members are: TIB, ZB Med, ZBW and Gesis (GER), The British Library (GB), TU Delft (NL), INIST (FRA), DTU (DEN), SND (SWE), ETH Library Zürich (CH) and CRUI (IT).

EPIC

**Mission**
The EPIC consortium has as its mission enabling partner data centers to register PIDs in any number for the data objects and collections they store independent of their possible later use for publications. This registration in general will be done by automated procedures using an API that also allows adding relevant information such as checksums, pointers to metadata and rights information, etc. EPIC is thus open for any European center that stores scientific/research data.

**Characteristics**
EPIC is setup as a highly reliable, persistent and high performance service. This is achieved through a network of strong data centers in Europe that share the same service, the same API and that have implemented a redundancy scheme. When one center is out of order the other centers will still resolve the PIDs. Currently 3 data centers share the service, but with the help of the European Union seven more strong data centers will join EPIC resulting in a broad European coverage. Registered PIDs will not be deleted as long as they are needed or the objects are accessible.

**Governance & Costs**
EPIC is controlled by its scientific user communities and organizations to ensure that it is devoted to the needs of the research community at large. This will also ensure that the overhead of the EPIC consortium will be small and restricted to essential services. Running the Handle Services based on tested software is not expensive, however, EPIC sees the need to establish help and support services that will require some funding. Currently and for the coming years funding is ensured. At a certain moment a low contribution will be required. The User Board will determine the funding structure.
**European PID Consortium (EPIC)**

EPIC is primarily meant to allow data centers in Europe to register the millions of data objects and collections that are being created as the result of the research process.

Current Partners:

- Computer center of the Max Planck Society and the University of Göttingen
- National Dutch Computer Center
- National Finnish Computer Center

Contact:
Ulrich Schwardmann  
GWDG - eResearch Group  
Am Fassberg 11  
Göttingen  
Germany  
Web: http://pidconsortium.eu  
Email: ulrich.schwardmann@gwdg.de

**DataCite**

DataCite is primarily meant to enhance the discoverability of data objects by running a central metadata repository of published data collections that can be cited in scientific publications.

DataCite is an official DOI-registration agency and has 16 members from 11 countries, working with more than 80 data centers.

Contact:
Jan Brase  
German National Library of Science and Technology (Managing Agent)  
Welfengarten 1B  
30167 Hannover  
Germany  
Web: http://datacite.org