DOI and DataCite

Establishing information infrastructures

Dr. Angelina Kraft

Digital object identifiers for research results: providers and best practice

Cluster K Workshop, 8 June 2015, Vienna

DataCite

FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for research data

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
Why? – Political significance!

- Social & political responsibility
- European Commission requirements
- Horizon 2020
- Open Access strategies
- Funding body requirements

OPEN ACCESS

RDA RESEARCH DATA ALLIANCE
Research Data Sharing without barriers

→Science policy requirement to publish research data

→Reusability of publicly funded research

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
Why? – Publishing companies!

- **STM Association – 2015 Report:**

  "…The explosion of data-intensive research is challenging publishers to create new solutions

  *to **link** publications to research data (…)*

  *to facilitate **data mining** and*

  *to **manage** the dataset as a potential unit of publication (…)*

Change continues to be rapid, with new leadership and coordination from

*the **R**esearch **D**ata **A**lliance (…)*
***research funders** have introduced or tightened (data) policies*
***data repositories** have grown in number and type (…) and*
***DataCite** was launched (…)*
*discovery services such as Thomson Reuters' **Data Citation Index**…"*

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
Why? – Publishing companies!

- **Brussels Declaration –**
  STM Association publishing companies

  *"**… Sets or sub-sets of data that are submitted with a paper** to a journal should wherever possible be made **freely accessible** to other scholars."*

- **Response: data journals**
  Example: Nature: *Scientific Data*

  "Scientific Data's central mission is to help foster the sharing and re-use of the data underpinning scientific research."

TIB | German National Library of Science and Technology

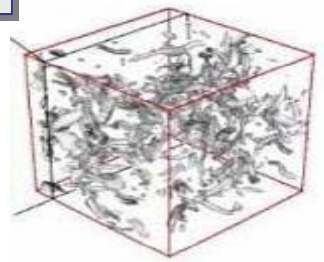
DataCite FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
Why? – Scientific significance!

- One thousand years ago,
  science was **empirical:**
  described natural phenomena

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

- Over the last one hundred years,
  a **theoretical** branch developed:
  building on models, generalisations



- In recent decades, an
  IT branch:
  Simulation of complex phenomena

- Today, science is
  data-based (eScience):
  Combination of theory, experiment & simulation

TIB | German National Library of Science and Technology

DataCite
FIND, ACCESS, AND REUSE DATA

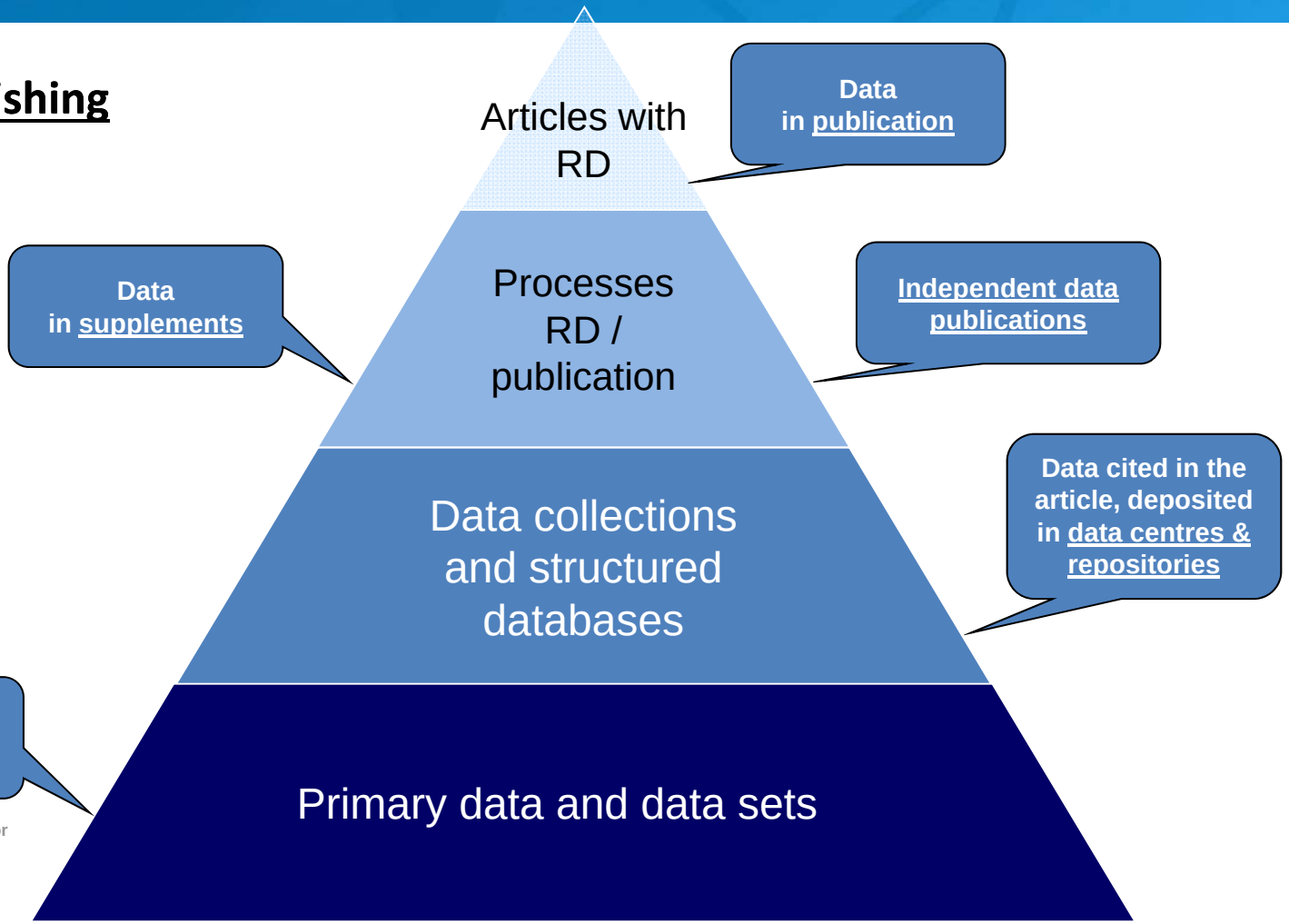
1. Persistent identification & DOI for data
But – scientific scepticism!

"A biologist would rather share their toothbrush than their gene name"

Mike Ashburner and others
Professor in Dept of Genetics, University of Cambridge, UK

1. Persistent identification & DOI for data
Data landscape – the theory

Options for publishing data:



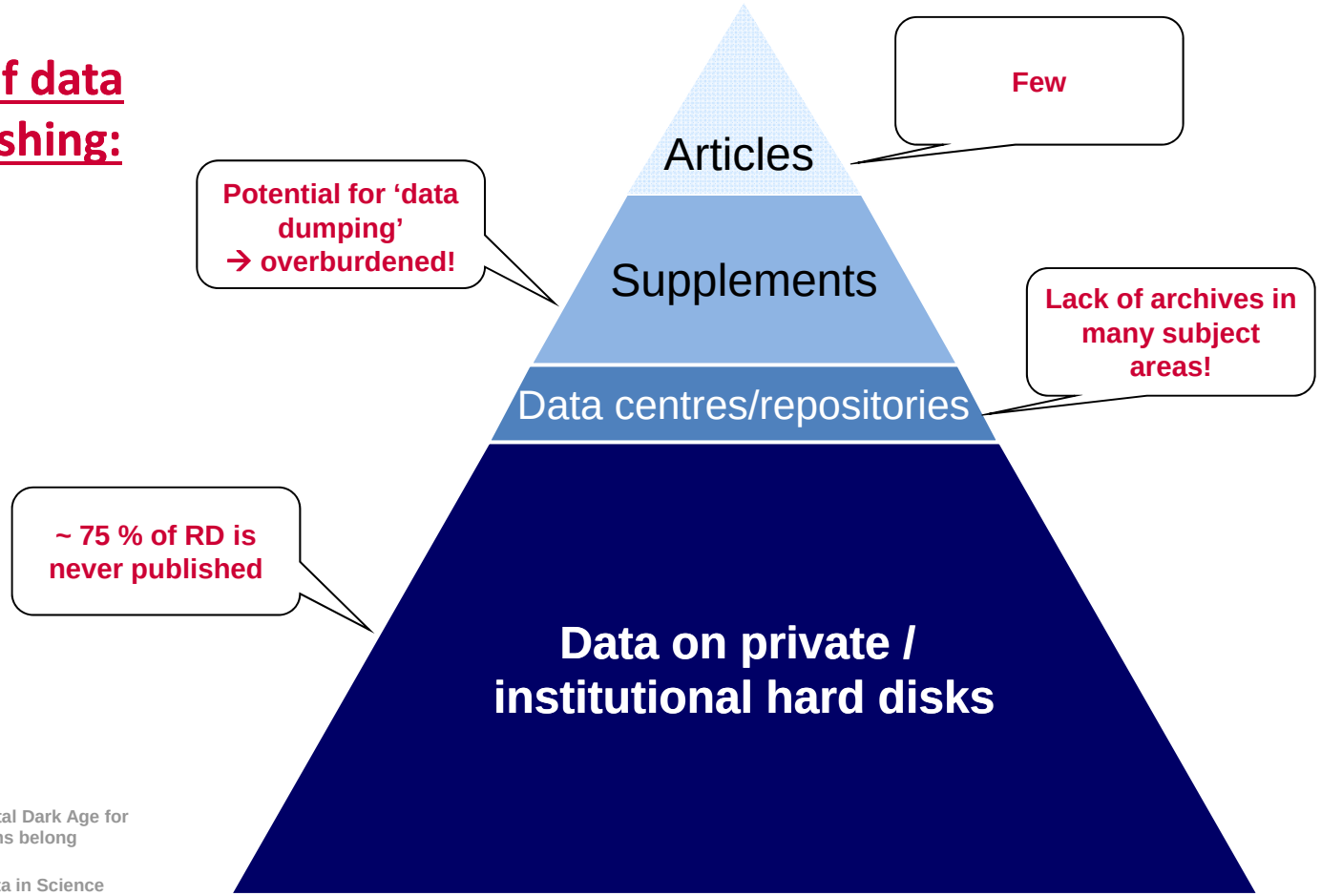
Modified based on
STM / Smit, E: Avoiding a Digital Dark Age for Data: why data and publications belong together
ICSTI workshop Delivering Data in Science
PARIS, 5 March 2012

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
Data landscape – the reality

Reality of data publishing:



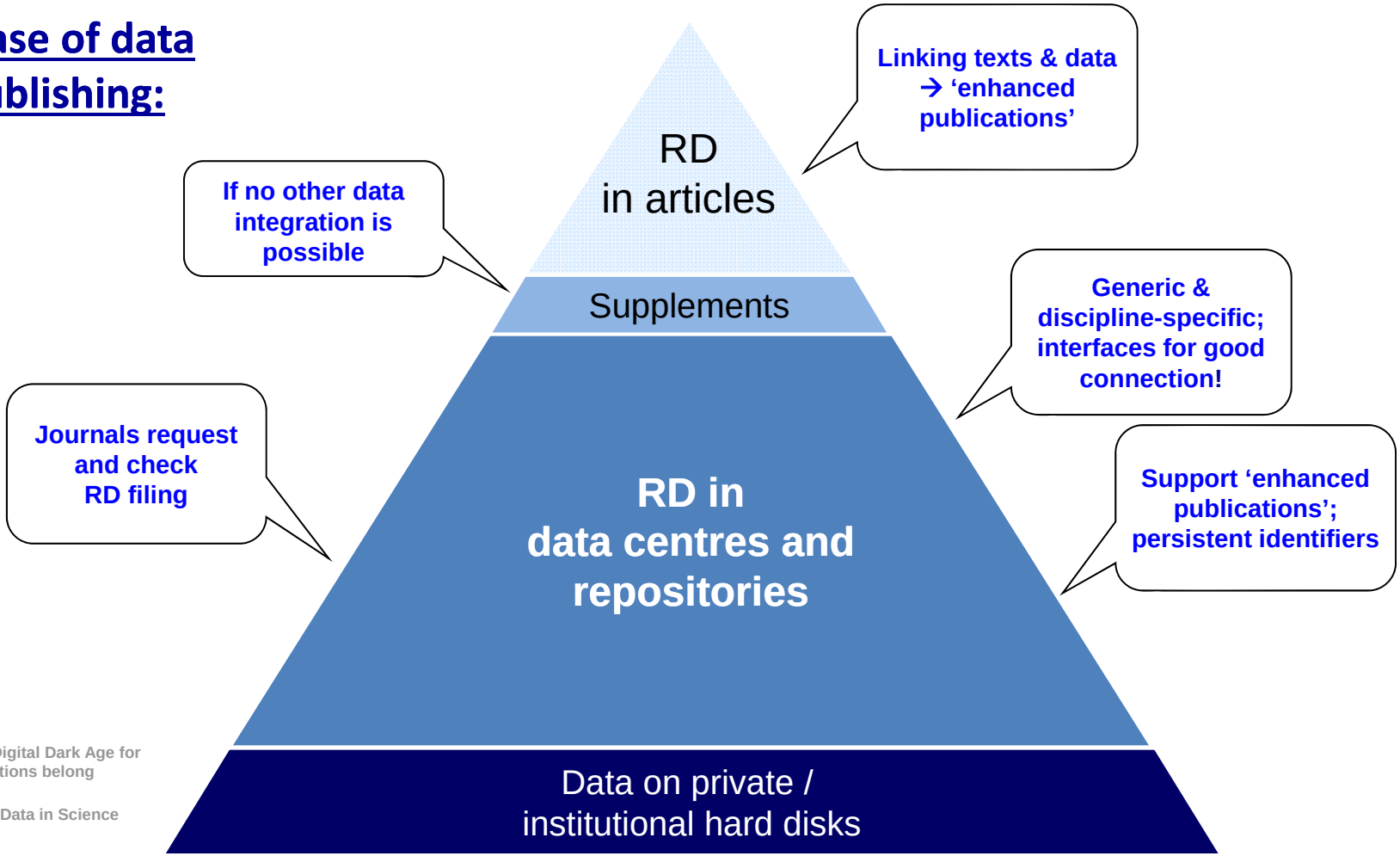
Modified based on
STM / Smit, E: Avoiding a Digital Dark Age for Data: why data and publications belong together
ICSTI workshop Delivering Data in Science
PARIS, 5 March 2012

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
Data landscape – the future?

Ideal case of data publishing:



Modified based on
STM / Smit, E: Avoiding a Digital Dark Age for Data: why data and publications belong together
ICSTI workshop Delivering Data in Science
PARIS, 5 March 2012

TIB | German National Library of Science and Technology

DataCite
FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
Advantages

- Clear referencing and citability
- Links data to other publications
- Increased visibility & enhanced access
- Transparent research
- Avoids duplication
- Promotes scientific cooperation
- Motivation for new research

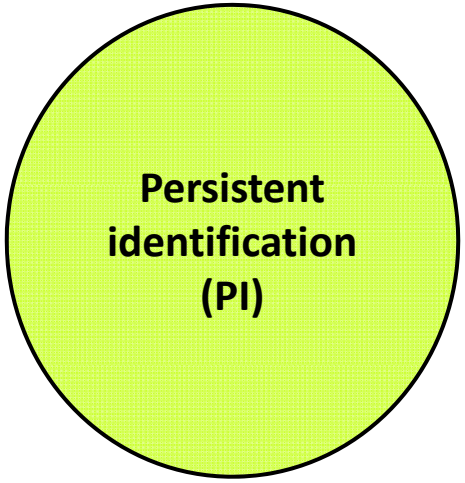
TIB | German National Library of Science and Technology

DataCite
FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
Properties

- Resource can be **clearly referenced & cited**
- **Persistent**, i.e. also beyond the life span of the identified object, if necessary
- Clear separation between identification of the resource and the location reference

- PI is undertaken by **registration agencies**:
  - Standards for structure and syntax
  - Resolving mechanism



TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

1. Persistent identification & DOI for data
DOI system

- International DOI Foundation (IDF) founded in 1998
- **Long-term persistence & accessibility to objects**
- Technology based on the Handle system.
- May 2012: DOI System ISO Standard 26324 was published
- Guaranteed, trustworthy responsibilities, uniform standards & work flows
- Quality control: obligatory metadata for each object
- IDF currently consists of nine registration agencies (RA)
- RA responsible for PI allocation and maintenance

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

2. DataCite

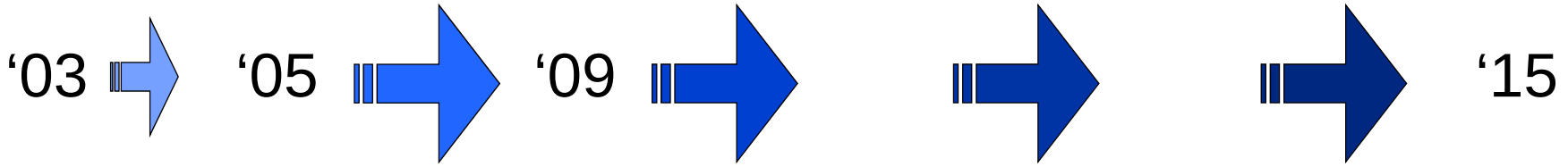
2. DataCite
Background

- **Global consortium** supported by local institutions
- Goal: **Publication infrastructure for data & non-textual content**
- Service provider for data centres/content providers
- Non-commercial, non-profit
- Standards, work flows and best practice
- Based on the DOI system

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

2. DataCite
Development



DFG-funded project with German World Data Centres

TIB allocates the first DOIs for data sets

Paris Memorandum.

DataCite is founded in London. Seven members.

Sturdy technical infrastructure.

Annual Meetings:

Hannover 2010
Berkeley 2011
Copenhagen 2012
Washington 2013
Nancy 2014
Paris 2015

25 members

8 associated members

19 countries

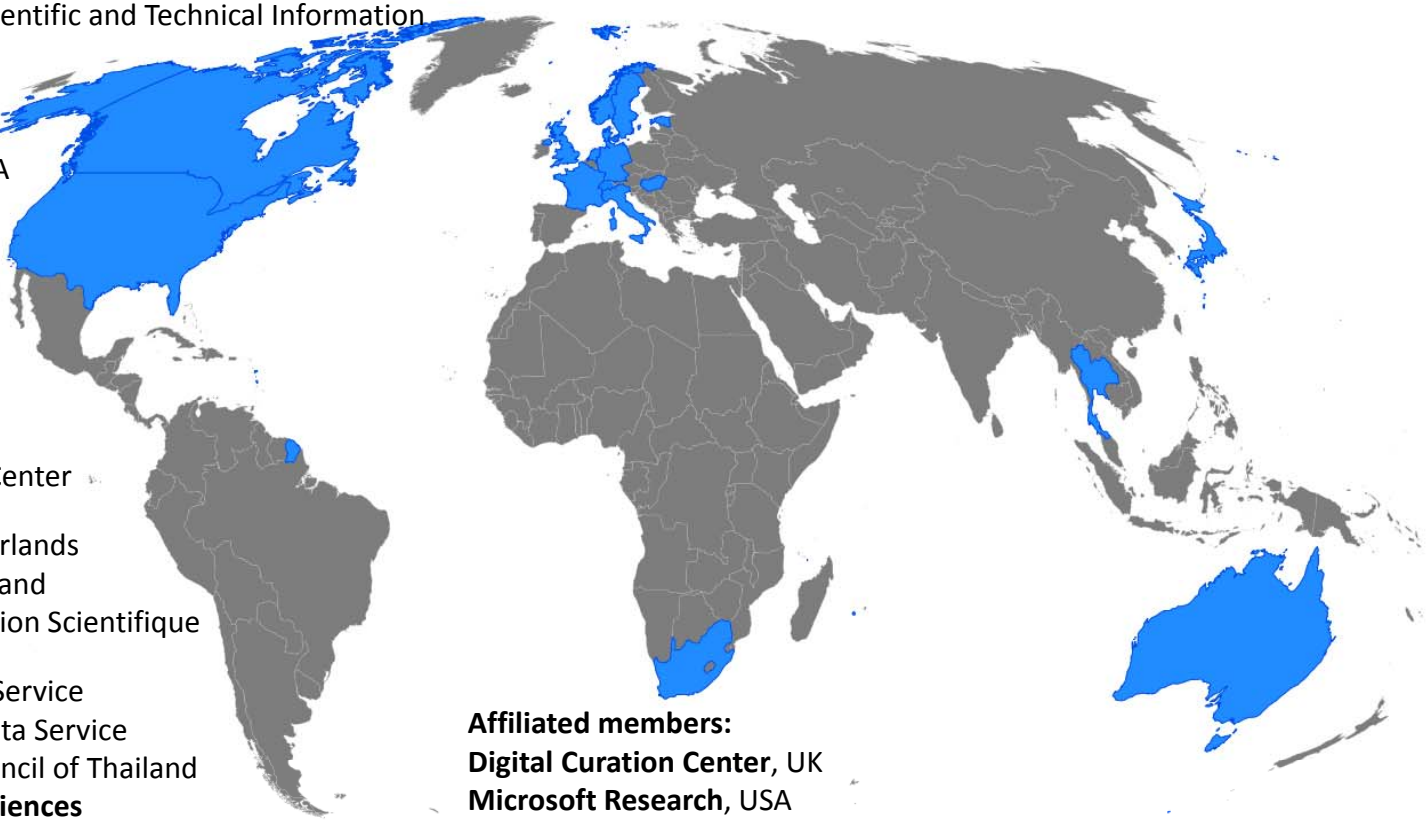
Over 5 million DOI names

TIB | German National Library of Science and Technology

DataCite
FIND, ACCESS, AND REUSE DATA

2. DataCite
Members

CISTI – Canada Institute for Scientific and Technical Information
California Digital Library, USA
Purdue University, USA
OSTI – Office of Scientific and Technical Information, USA
The British Library
TIB, Germany
ZB MED, Germany
ZBW, Germany
GESIS, Germany
University of Tartu, Estonia
JaLC – Japan Link Center
DTIC – Technical Information Center of Denmark
Library of TU Delft, The Netherlands
Library of ETH Zürich, Switzerland
INIST – L'Institut de l'Information Scientifique et Technique, France
SND – Swedish National Data Service
ANDS – Australian National Data Service
NRCT – National Research Council of Thailand
The Hungarian Academy of Sciences
CRUI – Conferenza dei Rettori delle Università Italiane
SAEON – South African Environmental Observation Network
CERN – European Organization for Nuclear Research
BIBSYS – Library System, Norway

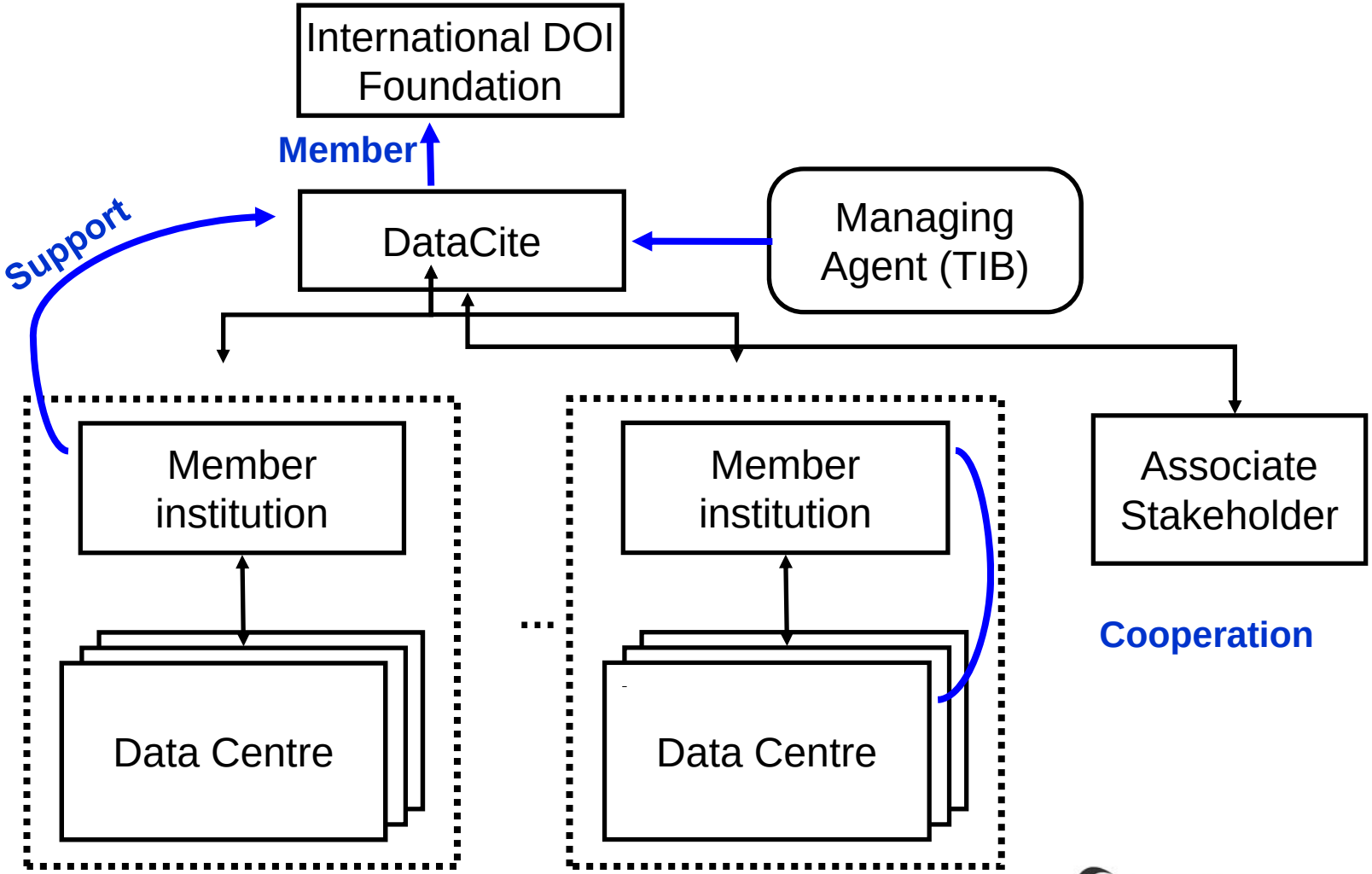


Affiliated members:
Digital Curation Center, UK
Microsoft Research, USA
ICPSR – Interuniversity Consortium for Political and Social Research, USA
KISTI – Korea Institute of Science and Technology Information
BGI – Bejiing Genomic Institute, China
IEEE, USA
Harvard University Library, USA
GWDG, Germany

TIB | German National Library of Science and Technology

DataCite
FIND, ACCESS, AND REUSE DATA

2. DataCite
Structure



TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

2. DataCite
Services

- Members and associated members:

  →Libraries, information and data centres

- Working Groups:

  → Metadata

  → Best practices

- Other services:

  → Metadata Store, Search, Stats, OAI Provider

  http://www.datacite.org/services

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

2. DataCite
Cooperative activities - I

- In cooperation with CrossRef:
  - http://crosscite.org/citeproc/
    Citation Formatter makes available citations in over 100 formats
  - http://crosscite.org/cn/
    Content Negotiation can be used to automatically obtain access to the (previously deposited) media formats of an object

- With STM Association publishing companies:
  - Improved ability to access & find research data
  - Promotion of bidirectional links between data sets & publications in data archives
  - Enhanced visibility of links between publications & data sets

TIB German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

2. DataCite
Cooperative activities - II

- Thomson Reuters - Data Citation Index
  - Harvesting metadata via DataCite
  - Advantages for customers:
    Access to DCI statistics

- THOR project
  - *Technical and Human infrastructure for Open Research*
  - *30 month project funded by the European Commission under the Horizon 2020*
  - Seamless integration between articles, data, and researchers

TIB | German National Library of Science and Technology

DataCite
FIND, ACCESS, AND REUSE DATA

2. DataCite
Cooperative activities - III

- re3data & DataBib
  - To merge and act under the auspices of DataCite as re3data

- MoU with RDA:
  - DataCite will become an "organisational member"

- Endorsement of the Force11 "Joint Declaration of Data Citation Principles"

TIB | German National Library of Science and Technology

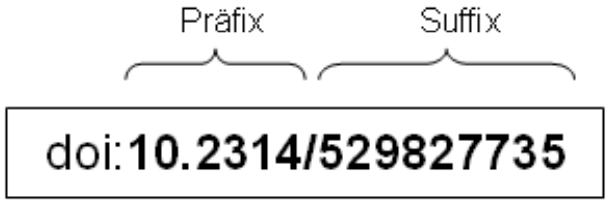
3. DOI registration

3. DOI registration
Types of content

- By 12/2014, over 4,250,000 DOI names had been allocated by DataCite for:
  - Research data (~45%)
  - Grey literature objects (~40%)
  - Images (~10%)
  - Medical case studies
  - Videos
  - Maps
  - Learning objects
  - Status in May 2015: 5,392337 DOI names

3. DOI registration
Demands placed on data centres

- Securing persistence
- Providing metadata & landing pages
- Securing data granularity (worthy of citation?)
- DOI syntax:



  - Prefix is allocated by DataCite
  - Suffix can be defined by the data centre
  - Clear string
  - Positive list: A-Z a-z 0-9 . : - _ /

→ New DOIs are resolvable after around 5 minutes
→ DOI update globally available after a max. of 24 hours

3. DOI registration
DataCite metadata schema - mandatory fields

- Identifier *(with type attribute)*
- Creator *(with type and name identifier attributes)*
- Title *(with optional type attribute)*
- Publisher
- Publication year

- Recommended citation:

  Creator (Publication Year): Title. Publisher. Identifier

3. DOI registration
DataCite metadata schema – optional/recommended fields

- **Subject** *(with scheme attribute)*
- **Contributor** *(with type and name identifier attributes)*
- **Date** *(with type attribute)*
- Language
- **Resource type** *(with description attribute)*
- Alternate identifier *(with type attribute)*
- **Related identifier** *(with type and relation type attributes)*
- Size
- Format
- Version
- Rights
- **Description** *(with type attribute)*
- **GeoLocation** *(with point, box and place)*

3. DOI registration
Citing with DOI - I - papers & research data

This is how for, example, the data set:

Kuhlmann, H et al. (2009):

Age models, iron intensity, magnetic susceptibility records and dry bulk density of sediment cores from around the Canary Islands.

doi:10.1594/PANGAEA.727522

is analysed in the following article:

Kuhlmann et al. (2004):

Reconstruction of paleoceanography off NW Africa during the last 40,000 years: influence of local and regional factors on sediment accumulation.

Marine Geology, 207(1-4), 209-224,

doi:10.1016/j.margeo.2004.03.017

TIB | German National Library of Science and Technology

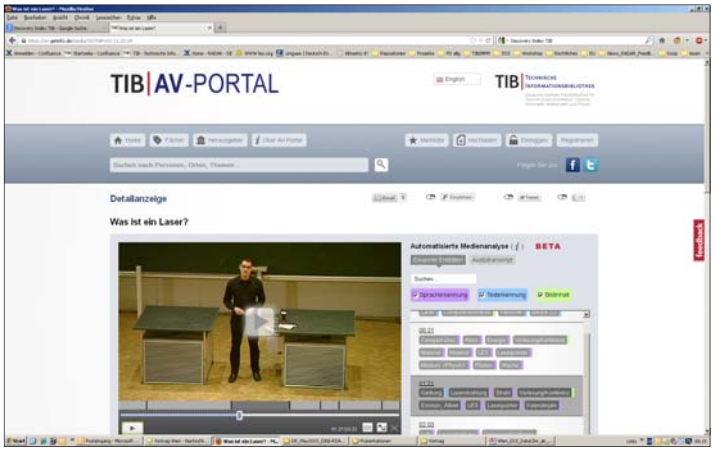
DataCite FIND, ACCESS, AND REUSE DATA

3. DOI registration
Citing with DOI - III – media fragment identifier

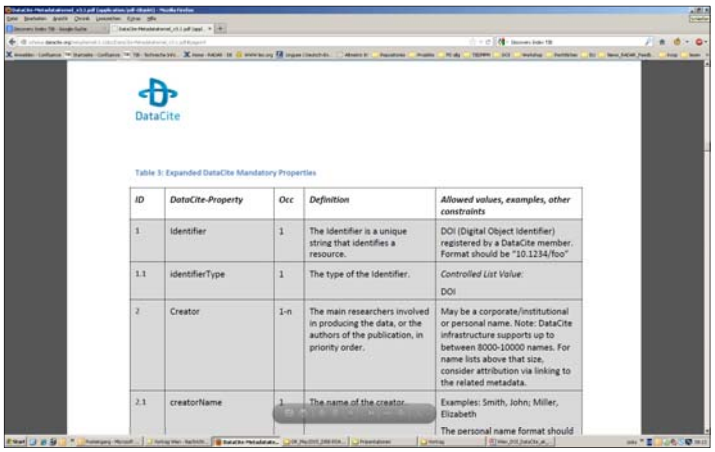
- Very precise citation of videos:

resolver DOI MFID

http://dx.doi.org/10.5446/393#t=01:21,02:04



- Can also be used for other media if fragmentation is supported:
  - PDF: doi.org/10.5438/0010#page=9



TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

3. DOI registration
Data granularity

Constantly being debated, but:

no universally valid guidelines (so far) for the granularity of research data!

Every object that is to be cited may be allocated a DOI!

3. DOI registration
DOI facts

- DOIs cannot be deleted
- A DOI should always persistently identify precisely one object
- A DOI refers to a landing page – this is where metadata & information about the object is noted
- Should the object identified by the DOI no longer be available, this has to be specified on the landing page

3. DOI registration
DataCite Metadatastore (MDS)

https://mds.datacite.org/

- Register a data set
- Update a data set
- Upload a metadata file
- Find a specific DOI

Individual operations

→ **User Interface (UI)**

- Register several data sets
- Update several data sets
- Upload several metadata files
- Retrieve metadata

"Bulk" operations

→ **Application Programming Interface (API)**

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA

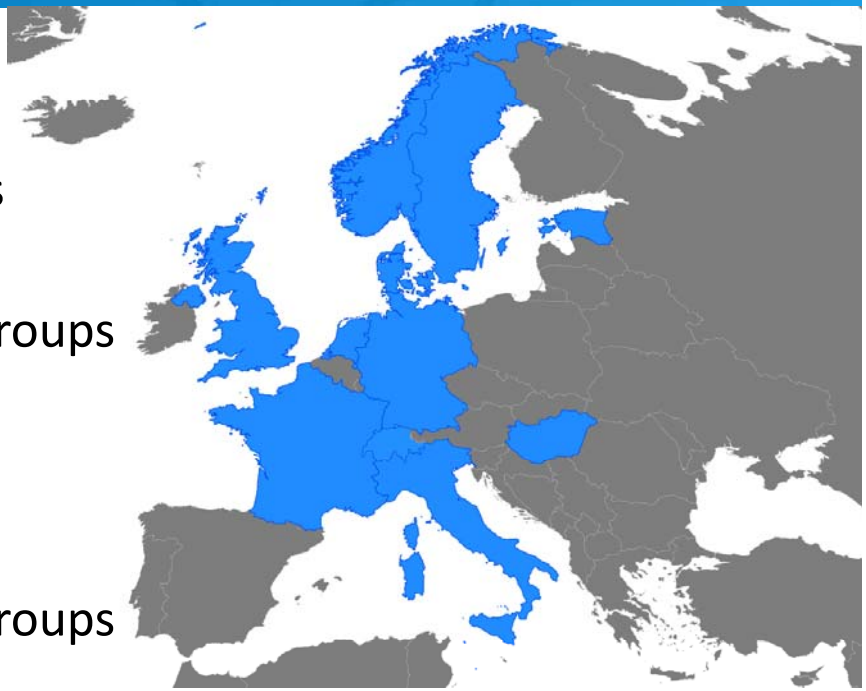
3. DOI registration
DataCite MDS - test environment

DataCite provides its own test environment in which all services can be tested in a closed system:
http://test.datacite.org

Resolver for test DOIs: http://dx.test.datacite.org

4. How to take part

4. How to take part
Possibilities



- **Membership**
  - Collaboration with local data centres
  - Registration of DOIs
  - Collaboration in DataCite Working Groups
  - Co-determination in DataCite
- **Associated membership**
  - Collaboration in DataCite Working Groups
  - Provision of advice for DataCite
- Cooperation with a member as a **data centre**
  - DOI registration for your data sets

TIB | German National Library of Science and Technology

DataCite FIND, ACCESS, AND REUSE DATA