

DOI und DataCite

Etablierung von Informationsinfrastrukturen

Dr. Angelina Kraft
Digital Object Identifier für Forschungsergebnisse: Anbieter und Best Practice
Cluster K Workshop, 08.06.2015, Wien



Übersicht

1. Persistente Identifizierung & DOI für Forschungsdaten
2. DataCite
3. Die DOI-Registrierung
4. How to take part
5. TIB-Beitrag Informationsinfrastruktur:
DOI-Service & RADAR-Projekt

1. Persistente Identifizierung & DOI für Forschungsdaten

1. Persistente Identifizierung & DOI für Daten

Warum? – Politische Bedeutung!

- Gesellschaftliche & politische Verantwortung
- Vorgabe der EU-Kommission
- Horizon 2020
- Open Access Strategien
- Anforderungen der Fördergeber



Research Data Sharing
without barriers

→ **Wissenschaftspolitische Anforderung nach Publikation von Forschungsdaten**

→ **Nachnutzbarkeit öffentlich geförderter Forschung**

1. Persistente Identifizierung & DOI für Daten

Warum? – Verlage!

- **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 **Research Data Alliance** (...)*

***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**...”*



1. Persistente Identifizierung & DOI für Daten

Warum? – Verlage!

- **Brüsseler Erklärung –
Verlage der STM Association**

„... 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.“



- **Antwort: Datenjournals
Beispiel: Nature: *Scientific Data***

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



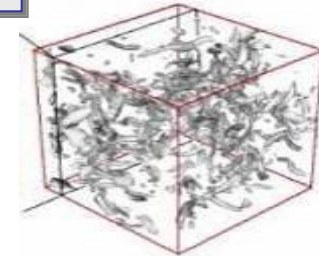
1. Persistente Identifizierung & DOI für Daten

Warum? – Wissenschaftliche Bedeutung!

- Vor eintausend Jahren war die Wissenschaft **empirisch**:
beschrieb Naturphänomene
- In den letzten einhundert Jahren entstand ein **Theoretischer** Zweig:
aufbauend auf Modellen, Generalisierungen
- In den letzten Dekaden ein **Informatischer** Zweig:
Simulation komplexer Phänomene
- Heute ist Wissenschaft **Datenbasiert** (eScience):
Vereinigung von Theorie, Experiment & Simulation



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



1. Persistente Identifizierung & DOI für Daten Aber – Wissenschaftliche Skepsis!

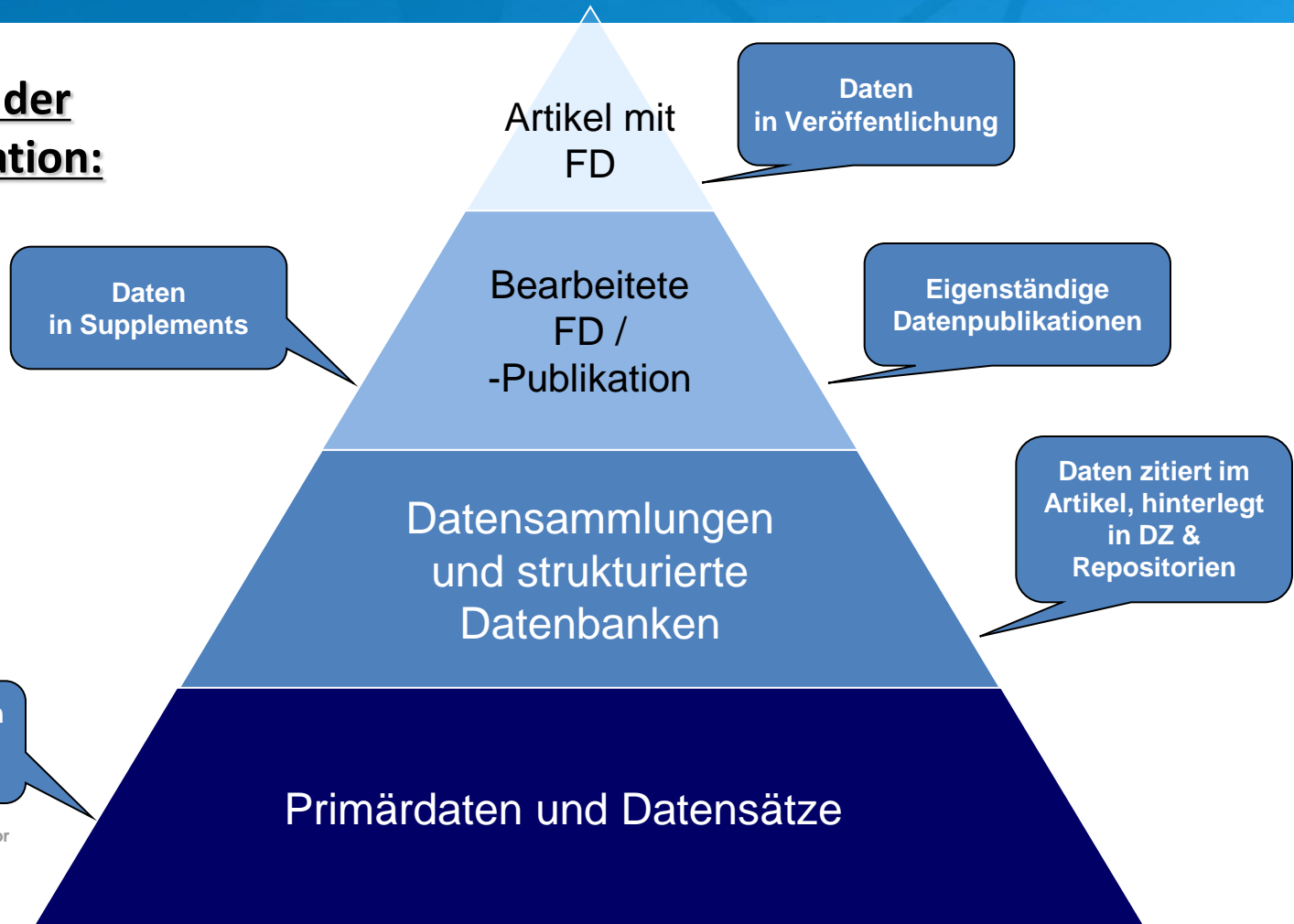
“A biologist would rather share their toothbrush than their gene name”

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



1. Persistente Identifizierung & DOI für Daten Datenlandschaft – Die Theorie

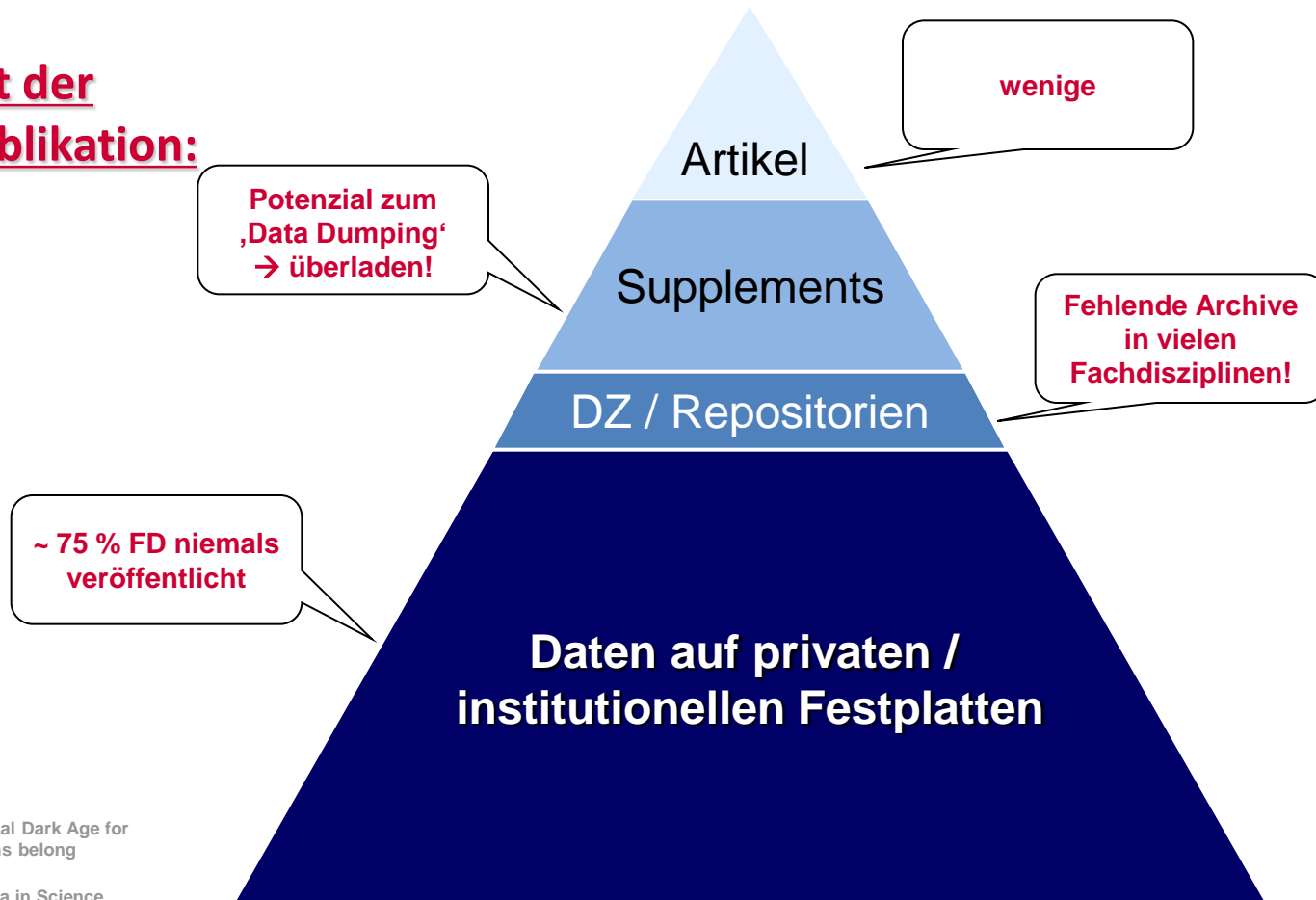
Möglichkeiten der Datenpublikation:



Modifiziert nach
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

1. Persistente Identifizierung & DOI für Daten Datenlandschaft – Die Realität

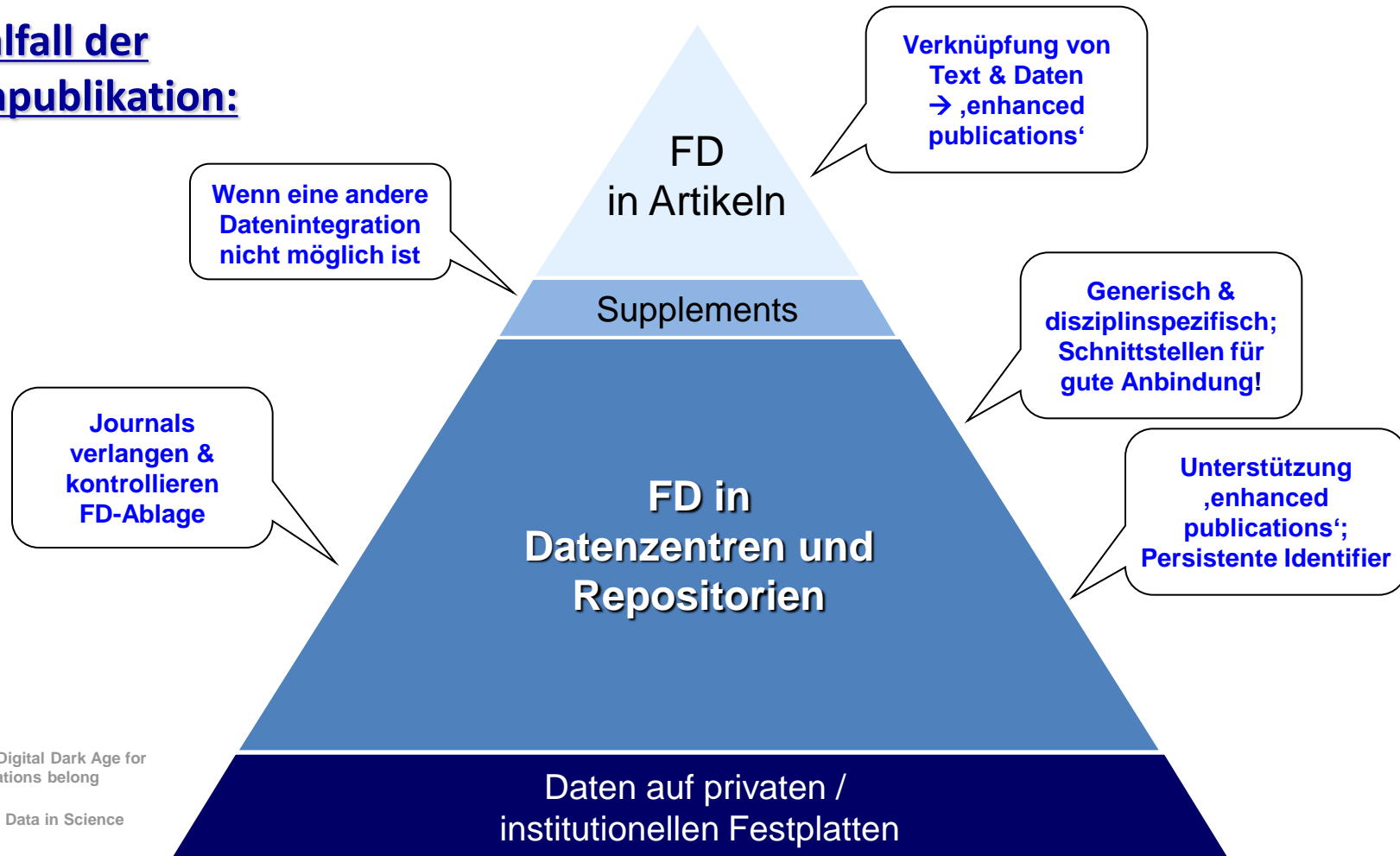
Realität der Datenpublikation:



Modifiziert nach
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

1. Persistente Identifizierung & DOI für Daten Datenlandschaft – Die Zukunft?

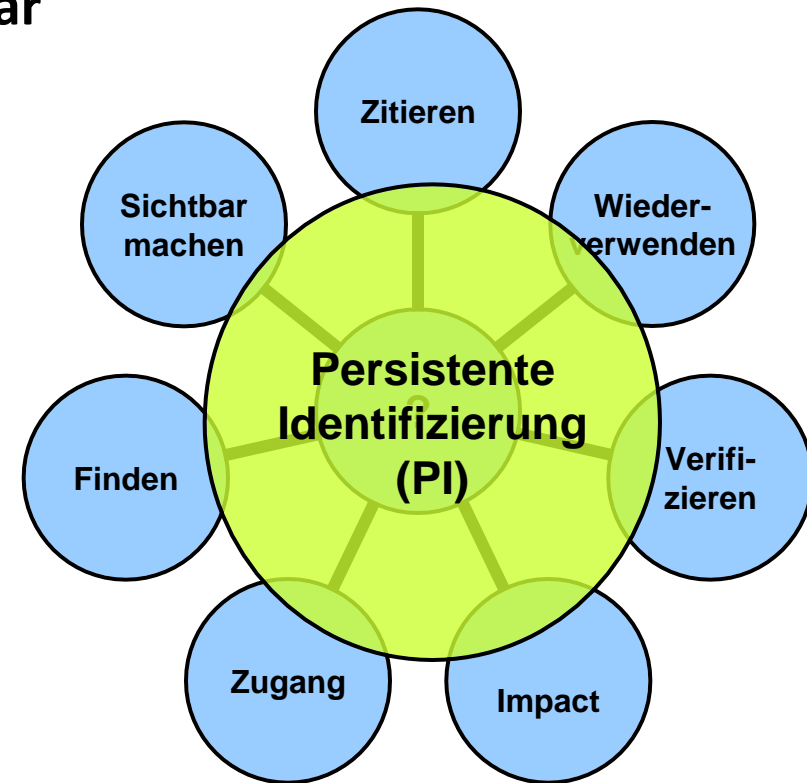
Idealfall der Datenpublikation:



Modifiziert nach
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

1. Persistente Identifizierung & DOI für Daten Eigenschaften

- Ressource wird **eindeutig referenzierbar & zitierfähig**
- **Dauerhaft**, d.h. ggf. auch über die Lebensdauer des identifizierten Objektes hinaus
- Klare Trennung von Identifikation der Ressource und der Standortreferenz
- PI wird von **Registrierungsagenturen** übernommen:
 - Standards für Struktur und Syntax
 - Resolving-Mechanismus



1. Persistente Identifizierung & DOI für Daten

DOI System

- International DOI Foundation (IDF) 1998 gegründet
- **Langfristige Persistenz & Zugänglichkeit zu Objekten**
- Basiert technisch auf dem Handle System
- Mai 2012: DOI System ISO Standard 26324 wurde publiziert
- Garantierte, vertrauenswürdige Verantwortlichkeiten, einheitliche Standards & Workflows

- Qualitätskontrolle: Obligatorische Metadaten für jedes Objekt
- IDF besteht derzeit aus 9 Registrierungsagenturen (RA)
- RA zuständig für PI-Vergabe und Pflege



DOI®, DOI.ORG® and shortDOI® sind Markennamen der International DOI Foundation

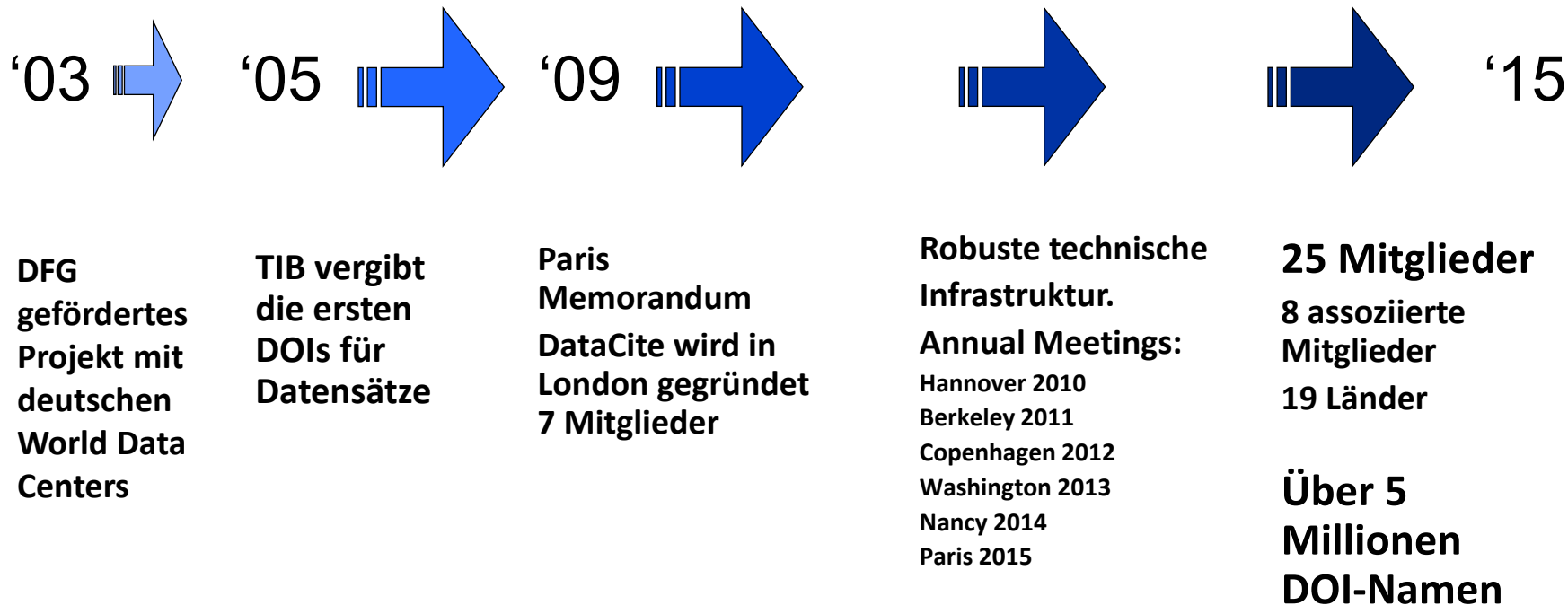
2. DataCite

2. DataCite Hintergrund

- **Globales Konsortium** getragen von lokalen Institutionen
- Ziel: **Publikationsinfrastruktur für Daten & nicht-textuelle Inhalte**
- Dienstleistungsanbieter für Datenzentren/Inhaltsanbieter
- Nicht-kommerziell, non-profit
- Standards, Workflows und Best Practice
- Basiert auf dem DOI System



2. DataCite Entwicklung



2. DataCite Mitglieder

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

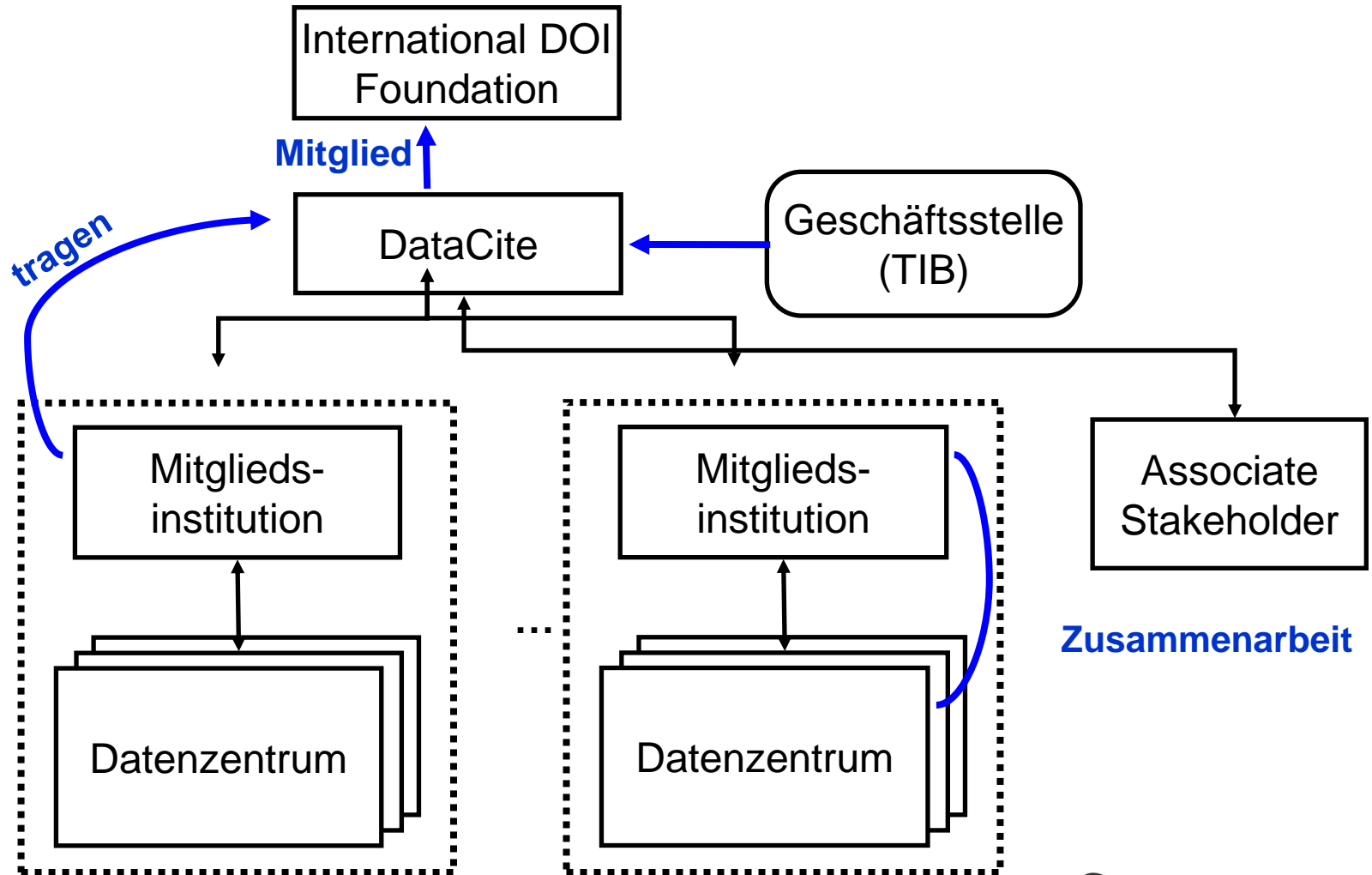
BGI – Beijing Genomic Institute, China

IEEE, USA

Harvard University Library, USA

GWDG, Germany

2. DataCite Struktur



2. DataCite Services

- Mitglieder und assoziierte Mitglieder:
 - Bibliotheken, Informations- und Datenzentren
- Working Groups:
 - Metadaten
 - Best Practices
- Weitere Services:
 - Metadata Store, Search, Stats, OAI Provider

<http://www.datacite.org/services>

2. DataCite Kooperationen - I

- In Zusammenarbeit mit CrossRef:



- <http://crosscite.org/citeproc/>

Citation Formater stellt Zitierungen in über 100 Formaten bereit

- <http://crosscite.org/cn/>

Über Content Negotiation kann maschinell auf (vorher hinterlegte) Medienformate eines Objektes zugegriffen werden

- Mit Verlagen der STM Association:



- Verbesserung Zugriff auf & Auffindbarkeit von Forschungsdaten
- Förderung von bidirektionalen Verlinkungen zw. Datensätzen & Publikationen in Datenarchiven
- Erhöhung der Sichtbarkeit von Links zw. Publikationen & Datensätzen

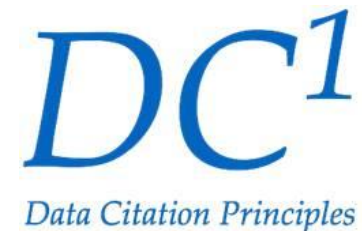
2. DataCite Kooperationen - II

- Thomson Reuters - Data Citation Index
 - Harvesten von Metadaten über DataCite
 - Vorteile für Kunden:
Zugriff auf die Statistiken des DCI
- ORCID – ODIN Projekt
 - *ORCID and DataCite Interoperability Network*
 - Aufnahme von Datensätzen in Publikationslisten
 - Nachverfolgung von Nutzung von Datensätzen
 - Verlinkung von Datensätzen mit zugehörigen Artikeln, Lizenzen und allen beteiligten Personen
 - Nachfolgeprojekt: THOR ab Juni 2015



2. DataCite Kooperationen - III

- re3data & DataBib
 - fusionieren und gehen als re3data unter die Schirmherrschaft von DataCite
- MoU mit RDA:
 - DataCite wird ein “organizational member”
- Endorsement of the Force11
„Joint Declaration of Data Citation Principles“



3. DOI-Registrierung

3. DOI-Registrierung

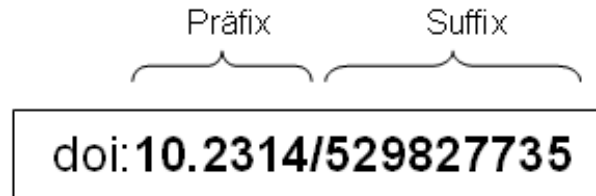
Inhaltstypen

- Bis 12/2014 wurden von DataCite über 4.250.000 DOI-Namen vergeben für:
 - Forschungsdaten (~45%)
 - graue Literaturobjekte (~40%)
 - Bilder (~10%)
 - Medizinische Fallstudien
 - Videos
 - Landkarten
 - Lernobjekte
 - Stand Mai 2015: 5.392.337 DOI-Namen

3. DOI-Registrierung

Anforderungen an Datenzentren

- Sicherstellung der Persistenz
- Bereitstellung von Metadaten & Landing Pages
- Sicherstellung der Datengranularität (zitierwürdig?)
- DOI-Syntax:



- Präfix wird von DataCite zugewiesen
- Suffix kann das Datenzentrum selber festlegen
- Eindeutiger String
- Positivliste: A-Z a-z 0-9 . : - _ /

→ Neue DOI sind nach etwa 5 Minuten auflösbar

→ DOI-Update nach max. 24 Stunden weltweit verfügbar

3. DOI-Registrierung

DataCite Metadatenchema - Pflichtfelder

- Identifier (*mit type Attribut*)
 - Creator (*mit type und nameIdentifier Attributen*)
 - Title (*mit optionalem type Attribut*)
 - Publisher
 - PublicationYear
-
- Zitierempfehlung:
Creator (PublicationYear): Title. Publisher. Identifier

3. DOI-Registrierung

DataCite Metadatenchema – Opt. & Empfohlene Felder

- **Subject** *(mit scheme Attribut)*
- **Contributor** *(mit type und nameIdentifier Attributen)*
- **Date** *(mit type Attribut)*
- Language
- **ResourceType** *(mit description Attribut)*
- Alternateldentifizier *(mit type Attribut)*
- **RelatedIdentifier** *(mit type und relationType Attributen)*
- Size
- Format
- Version
- Rights
- **Description** *(mit type Attribut)*
- **GeoLocation** *(mit point, box und place)*

3. DOI-Registrierung Zitierung mit DOI - I - Paper & Forschungsdaten

So wird beispielsweise der Datensatz:

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](https://doi.org/10.1594/PANGAEA.727522)

PANGAEA
Data Publisher for Earth & Environmental Science

Data Description

Citation: Kuhlmann, H et al. (2004) Age models, iron intensity, magnetic susceptibility records and dry bulk density of sediment cores from around the Canary Islands. doi:10.1594/PANGAEA.727522.
Supplement to: Kuhlmann, Holger; Freudenthal, Tim; Helmke, Peer; Meggers, Helge (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

Abstract: A set of 43 sediment cores from around the Canary Islands is used to characterise this region, which intersects meridional climatic regimes and zonal productivity gradients in a high spatial resolution. Using rapid and nondestructive core logging techniques we carried out Fe intensity and magnetic susceptibility (MS) measurements and created a stack on the basis of five stratigraphic reference cores, for which a stratigraphic age model was available from AMS and ¹⁴C analyses on planktonic foraminifera. By correlation of the stack with the Fe and MS records of the other cores, we were able to develop age depth models at all investigated sites of the region. We present the bulk sediment accumulation rates (ASR) of the Canary Islands region as an indicator of shifts in the upwelling-influenced areas for the Holocene (0-12 ky), the deglaciation (12-18 ky) and the last glacial (18-40 ky). General observations are an enhanced productivity during glacial times with highest values during the deglaciation. The main differences between the analysed time intervals we interpret as result of the sea-level effects, changes in the extent of high productivity areas, and current intensity.

Project(s): Geosciences, University of Bremen (DeB) & Center for Marine Environmental Sciences (MARUM) &

Coverage: Median Latitude: 28.28307 ° Median Longitude: -13.948692 ° South-bound Latitude: -35.250000 ° West-bound Longitude: -78.000000 ° North-bound Latitude: 32.703000 ° East-bound Longitude: -10.288700 °
Date/Time Start: 1995-05-28T00:00:00 ° Date/Time End: 1999-10-19T08:04:00 °

Event(s): GeoB1346-2 & * Latitude: -35.250000 ° Longitude: -78.000000 ° Date/Time: 1995-05-29T00:00:00 ° Elevation: -4306.0 m ° Location: South-East Pacific ° * Campaign: SO1031 (CHPAP) ° * Basis: Sonne ° * Device: Multiple opening/closing net (MSN) ° * Comment: max. depth: 500 m
GeoB4205-2 & * Latitude: 32.180000 ° Longitude: -11.648293 ° Date/Time: 1996-12-07T13:42:00 ° Elevation: -3206.0 m ° Recovery: 5.91 m ° Location: Agadir Canyon ° * Campaign: M071 ° * Basis: Meteor (1898) ° * Device: Gravity corer (Kiel type) (SL) °
GeoB4206-1 & * Latitude: 31.488333 ° Longitude: -11.015000 ° Date/Time: 1996-12-07T20:10:00 ° Elevation: -1949.0 m ° Recovery: 5.71 m ° Location: Agadir Canyon ° * Campaign: M071 ° * Basis: Meteor (1898) ° * Device: Gravity corer (Kiel type) (SL) °

License: CC BY Creative Commons Attribution 3.0 Unported

Size: 87 datasets

Download Data: Download ZIP file containing all datasets: as tab-delimited text (use the following character encoding: [UTF-8 (Unicode PANGAEA files)] [2])

in folgendem Artikel analysiert:

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](https://doi.org/10.1016/j.margeo.2004.03.017)

ScienceDirect

Marine Geology
Volume 207, Issues 1-4, 30 June 2004, Pages 209-224

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

H. Kuhlmann, T. Freudenthal, P. Helmke, H. Meggers
doi:10.1016/j.margeo.2004.03.017

Abstract: A set of 43 sediment cores from around the Canary Islands is used to characterise this region, which intersects meridional climatic regimes and zonal productivity gradients in a high spatial resolution. Using rapid and nondestructive core logging techniques we carried out Fe intensity and magnetic susceptibility (MS) measurements and created a stack on the basis of five stratigraphic reference cores, for which a stratigraphic age model was available from AMS and ¹⁴C analyses on planktonic foraminifera. By correlation of the stack with the Fe and MS records of the other cores, we were able to develop age depth models at all investigated sites of the region. We present the bulk sediment accumulation rates (ASR) of the Canary Islands region as an indicator of shifts in the upwelling-influenced areas for the Holocene (0-12 ky), the deglaciation (12-18 ky) and the last glacial (18-40 ky). General observations are an enhanced productivity during glacial times with highest values during the deglaciation. The main differences between the analysed time intervals we interpret as result of the sea-level effects, changes in the extent of high productivity areas, and current intensity.

Keywords: Canary Islands, sediment accumulation rates, coastal upwelling, sea-level, Holocene, last glacial

3. DOI-Registrierung Zitierung mit DOI - II – Media Fragment Identifier

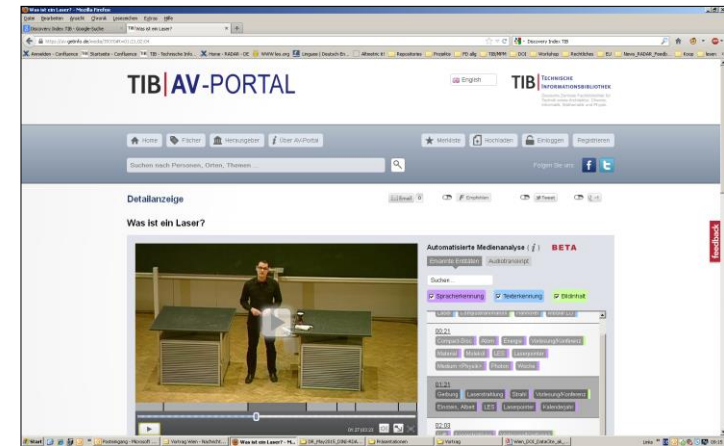
- Punktgenaue Zitierung von Videos:

resolver **DOI** **MFID**

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

- Auch für andere Medien einsetzbar wenn Fragmentierung unterstützt wird:

- PDF: doi.org/10.5438/0010#page=9



ID	DataCite-Property	Occ	Definition	Allowed values, examples, other constraints
1	Identifier	1	The identifier is a unique string that identifies a resource.	DOI (Digital Object Identifier) registered by a DataCite member. Format should be "10.1234/foo"
1.1	identifierType	1	The type of the identifier.	Controlled List Value: DOI
2	Creator	1-n	The main researchers involved in producing the data, or authors of the publication, in priority order.	May be a corporate/institutional or personal name. Note: DataCite infrastructure supports up to between 8000-10000 names. For name lists above that size, consider attribution via linking to the related metadata.
2.1	creatorName	1	The name of the creator.	Examples: Smith, John; Miller, Elizabeth The personal name format should

3. DOI-Registrierung Datengranularität

Immer wieder Thema, aber:

Keine allgemeingültigen Richtlinien (bisher) für die Granularität von Forschungsdaten!

Jedes Objekt, welches zitiert werden soll, kann einen DOI bekommen!

3. DOI-Registrierung

DOI-Fakten

- DOI können nicht gelöscht werden
- Ein DOI sollte immer genau ein Objekt dauerhaft identifizieren
- Ein DOI verweist auf eine Landing Page – dort sind Metadaten & Informationen zum Objekt vermerkt
- Sollte das Objekt, das durch den DOI identifiziert wird, nicht mehr verfügbar sein, muss dies auf der Landing Page angegeben werden

3. DOI-Registrierung DataCite Metadastore (MDS)

<https://mds.datacite.org/>

- Registrieren eines Datensatzes
- Aktualisierung eines Datensatzes
- Hochladen einer Metadaten-datei
- Finden eines bestimmten DOIs



Einzeloperationen

➔ **User Interface (UI)**



- Registrieren mehrerer Datensätze
- Aktualisierung mehrerer Datensätze
- Hochladen mehrerer Metadaten-dateien
- Abrufen von Metadaten



„Bulk“ Operationen

➔ **Application Programming Interface (API)**

3. DOI-Registrierung

DataCite MDS - Testumgebung

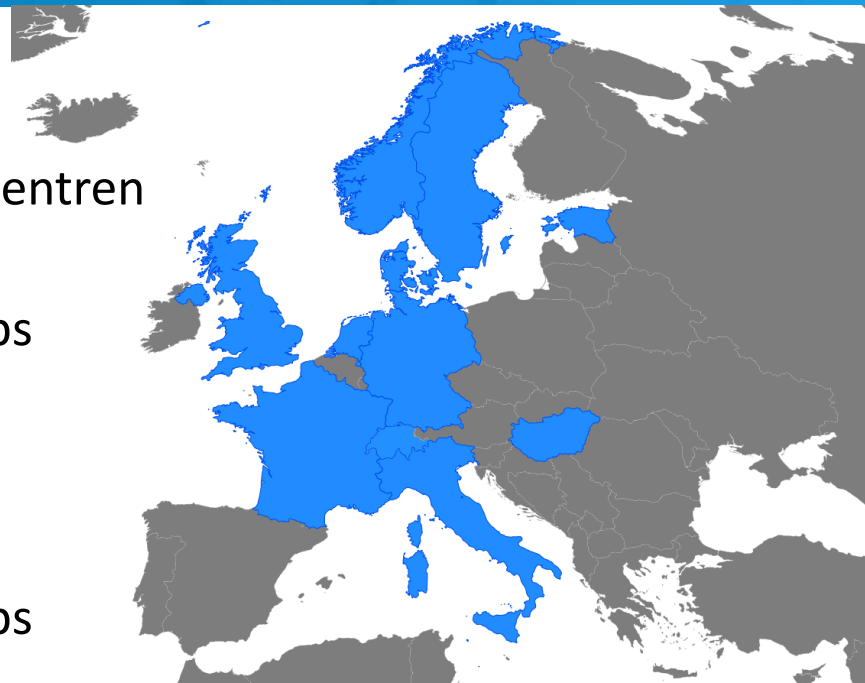
DataCite stellt eine eigene Test-Umgebung bereit, in der alle Services in einem abgeschlossenen System ausprobiert werden können: <http://test.datacite.org>

Resolver für Test-DOIs: <http://dx.test.datacite.org>

4. How to take part

4. How to take part Möglichkeiten

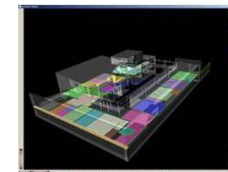
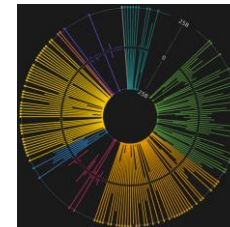
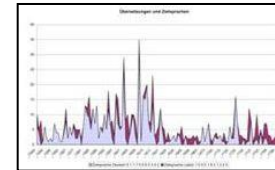
- **Mitgliedschaft**
 - Zusammenarbeit mit lokalen Datenzentren
 - Registrierung von DOIs
 - Mitarbeit in DataCite Working Groups
 - Mitbestimmung in DataCite
- **Assoziierte Mitgliedschaft**
 - Mitarbeit in DataCite Working Groups
 - Beratung für DataCite
- Zusammenarbeit mit einem Mitglied als **Datenzentrum**
 - DOI-Registrierung für Ihre Datensätze



5. TIB-Beitrag Informationsinfrastruktur: DOI-Service & RADAR-Projekt

5. TIB-Beitrag Informationsinfrastruktur DOI-Service - Scope

- TIB registriert Objekte aus dem Bereich Technik und Naturwissenschaften
 - Forschungsdaten
 - Graue Literatur
 - Open Access Artikel aus akademischen Einrichtungen
 - Videos
 - Bilder
 - Digitalisierte Kulturobjekte
 - Software
 - Weitere nicht-textuelle Objekte



- Zugang: Portal **GetInfo**
FIND THE WORLD OF
SCIENCE AND TECHNOLOGY

5. TIB-Beitrag Informationsinfrastruktur DOI-Service – Zahlen & Fakten

- DOI-Registrierungen (18.05.2015)
 - 1 021 912 DOI Namen (ca. 86 500 in 2015)
 - ~ 76% Forschungsdaten, 22% graue Lit., 0,25% AV-Medien
- 82 Datenzentren
 - Davon 3 aus Österreich:
 - European Society of Radiology + angegliederte Gesellschaften
 - JOANNEUM RESEARCH Forschungsgesellschaft mbH
 - Institute of Science and Technology Austria
 - 9 weitere aus dem Ausland

5. TIB-Beitrag Informationsinfrastruktur Wohin mit meinen (fachspezifischen) Daten?

Herausforderung 'Long Tail' Daten:

- Heterogen
- Fehlendes Anerkennungskonzept für die wissenschaftliche Leistung der Datenerzeugung & Publikation
- Kosten für Infrastrukturaufbau & nachhaltigen! Betrieb
- Gewährleistung Auffindbarkeit & Nutzbarkeit der Forschungsdaten

→ **Fachspezifisches Datenrepositorium**

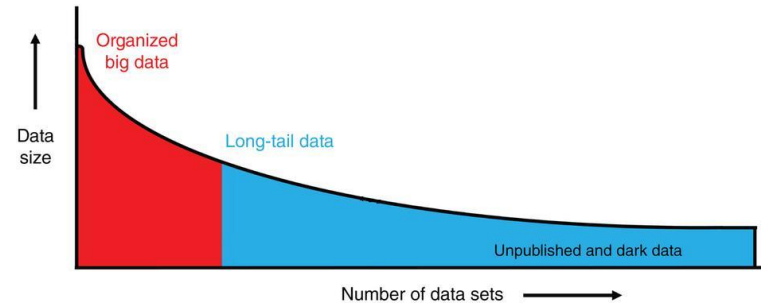
→ Übersicht: www.re3data.org

oder

→ **Disziplinübergreifendes Repositorium**

→ Generisches Repositorium: RADAR (De)

→ www.radar-projekt.org



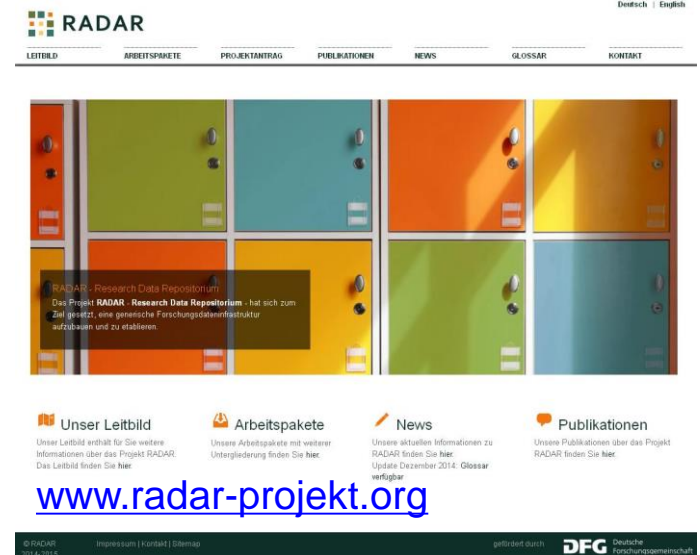
Source: Ferguson et al. (2014): Big data from small data: data-sharing in the 'long tail' of neuroscience. DOI: 10.1038/nn.3838

“The majority of datasets produced through research are part of the ‘Long Tail of Research Data’”

Source: Humphrey C (2014): OpenAIRE-COAR Conference, Athens

5. TIB-Beitrag Informationsinfrastruktur RADAR – Research Data Repository

- **Projektziel:**
Etablierung eines digitalen Datenrepositoriums (RADAR) als Basisdienstleistung für wissenschaftliche Institutionen zur Archivierung und Publikation von Forschungsdaten
- **Zweistufiger Ansatz:**
 - 1. Disziplinübergreifende Datenarchivierung**
5 – 10 – 15 Jahre, Handle
Abgestufte Zugriffsrechte
 - 2. Erweitertes Angebot Datenpublikation**
Adaptives Metadatenschema, DOI
Embargo & Peer-Review Support
- **Laufzeit:** 2013 – 2014 – 2015 – (2016)



➔ **Generisches End-Point
Repositorium mit
Dienstleistungen für
Wissenschaftler/Institutionen**

5. TIB-Beitrag Informationsinfrastruktur RADAR – Research Data Repository

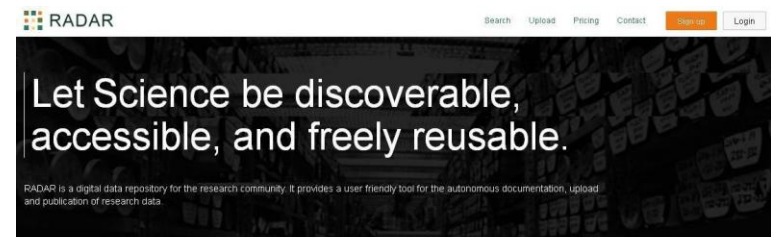
- **RADAR = Informationsinfrastruktur** für ‚heimatlose‘ Forschungsdaten
- Deutschlandweites **Verbundprojekt**
- Workshop 23.06.2015 – Frankfurt am Main



SCHRITTE ZUR ARCHIVIERUNG & PUBLIKATION



- 1 Registrierung / Anmeldung
- 2 Datenauswahl
- 3 Dienstleistungsmodelle:
 - A Archivierung
 - B Datenpublikation und -archivierung
- 4 Dateneingabe
- 5 Lizenzwahl
- 6 Datentransfer & Checksumming
- 7 Archivierung
- 8 Vergabe von persistenten Identifiern
- 9 Reporting an den Datengeber



Vielen Dank für Ihre Aufmerksamkeit!



