

Entwicklung von Standards und Best Practices im Bereich der Forschungsdatenpublikation

Ein Blick auf die Arbeit von DataCite

Open-Access-Tage 2015, Universität Zürich

Barbara Hirschmann, 8.9.2015

An aerial photograph of the ETH Zurich campus, showing various buildings, a large lake in the background, and distant mountains. A semi-transparent white rectangular box is overlaid on the center of the image, containing the text 'Wo stehen wir heute?'.

Wo stehen wir heute?

Wo stehen wir heute?



Quelle: Rietz, Helga: «Data Sharing» in der Wissenschaft. Die Daten der anderen. Neue Zürcher Zeitung, 22.7.2015.

Wo stehen wir heute?

Our first data set is the Bureau of Justice Statistics "Murder Cases in 33 Large Urban Counties." This is a random sample of homicide cases drawn from prosecutors' files. The data set includes information on offender characteristics, victim characteristics and trial outcomes for 2800 murders. The 75 largest counties account for more than half of the murders in the U.S. each year. This data set brings together information on the crime, the offender, the victim, and the sentence. Such information cannot all be linked in other larger data sets such as the Uniform Crime Reporting (UCR) Data or the National Crime Victimization Survey (NCVS). Most crime

Supplementary Material

Supplemental Data

[Click here to view.](#)

Acknowledgments

We thank Junghwa Seo, Lei Cho, and Jongmin Kim for technical assistance and Hyunjung Lim for consultation on image handling.

This work was supported, in whole or in part, by National Institutes of Health Grants AG5131 and AG18440. This work was also supported by the Disease Network Research Program (Grant 20090084180) from the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology, Republic of Korea and by the Korea Science and Engineering Foundation funded by the Korea government (Grant 200900000707).

The microarray data reported in this paper have been deposited to the Gene Expression Omnibus (GEO) data base under accession number [GSE11574](#).

S

The on-line version of this article (available at [http://www.jco.org](#)) contains supplemental files S1-S3 and Tables 1-4.

†A. Jang, H.-J. Lee, J.-E. Suk, J.-W. Jung, K.-P. Kim, and S.-J. Lee, submitted for publication.

Wo stehen wir heute?

- Forschungsdaten sind nach wie vor größtenteils nicht öffentlich zugänglich
- Forschungsdaten werden nicht oder uneinheitlich mit Metadaten beschrieben
- Für das Referenzieren und Zitieren von Forschungsdaten hat sich bisher kein Standard durchgesetzt

An aerial photograph of the ETH Zurich campus, showing various buildings, a large lake in the background, and distant mountains. A semi-transparent white rectangular box is overlaid on the center of the image, containing the text 'Wo möchten wir hin?' in a bold, black, sans-serif font.

Wo möchten wir hin?

Publikation in vertrauenswürdigen Repositorien

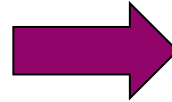
The image shows two overlapping screenshots of data repositories. The top screenshot is the Dryad website, which features a green logo and navigation links like 'About', 'For researchers', and 'Submit data now'. The bottom screenshot is the Dataverse website, showing a search interface for 'Harvard Dataverse' with various filters and search results.

- Vertrauenswürdiger Betreiber
- Sicherstellung des langfristigen Zugriffs
- Bereitstellung von Kontextinformationen (Metadaten)
- Persistente Identifizierung der Datensätze (z.B. via DOI)

Standardisierte Referenzen und Zitate

Informelle Referenzen

Our first data set is the Bureau of Justice Statistics "Murder Cases in 33 Large Urban Counties." This is a random sample of homicide cases drawn from prosecutors' files. The data set includes information on offender characteristics, victim characteristics and trial outcomes for 2800 murders. The 75 largest counties account for more than half of the murders in the U.S. each year. This data set brings together information on the crime, the offender, the victim, and the sentence. Such information cannot all be linked in other larger data sets such as the Uniform Crime Reporting (UCR) Data or the National Crime Victimization Survey (NCVS). Most crime



Formelle Referenzen

U.S. Dept. of Justice, Bureau of Justice Statistics (1996): Murder cases in 33 large urban counties in the United States, 1988. Version 1. Inter-university Consortium for Political and Social Research.
<http://dx.doi.org/10.3886/ICPSR09907.v1>

Supplementary Material

Supplemental Data
[Click here to view.](#)

Acknowledgments

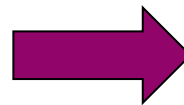
We thank Junghwa Seo, Lei Cho, and Jongmin Kim for technical assistance and Hyunjung Lim for consultation on image handling.

*This work was supported, in whole or in part, by National Institutes of Health Grants AG5131 and AG18440. This work was also supported by the Disease Network Research Program (Grant 20090084180) from the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology, Republic of Korea and by the Korea Science and Engineering Foundation funded by the Korea government (Grant 20090083737).

The microarray data reported in this paper have been deposited to the Gene Expression Omnibus (GEO) data base under accession number [GSE11574](#).

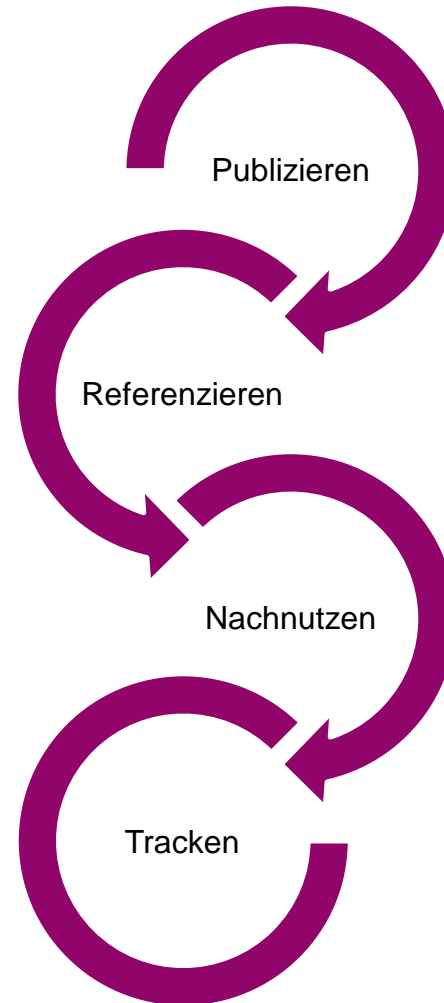
The on-line version of this article (available at <http://www.jco.org>) contains supplemental [Figs. S1-S8](#) and [Tables 1-4](#).

[†]A. Jang, H.-J. Lee, J.-E. Suk, J.-W. Jung, K.-P. Kim, and S.-J. Lee, submitted for publication.



Lee, Seung-Jae; Lee, He-Jin; Cho, Ji-Hoon; Rho, Sangchul; Hwang, Daehee (2008): GSE11574: The responses of astrocytes stimulated by extracellular a-synuclein. Gene Expression Omnibus.
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11574>

Wo möchten wir hin?



Wie kann DataCite dazu beitragen?

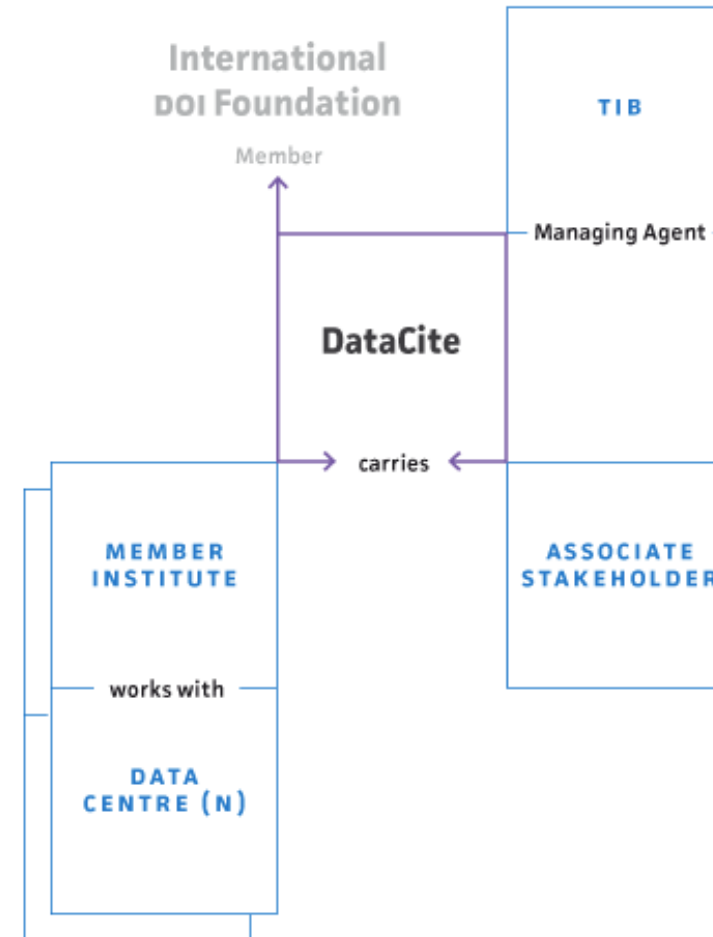
Das DOI-System

- **International DOI Foundation (IDF)**
 - 1998 gegründet
 - besteht derzeit aus 9 Registrierungsagenturen, darunter CrossRef und DataCite
- **DOI-System**
 - im Mai 2012 als ISO Standard 26324 publiziert
 - garantiert vertrauenswürdige Verantwortlichkeiten und einheitliche Standards
- **Registrierungsagenturen**
 - zuständig für DOI-Vergabe und -Pflege



DataCite: Organisation

- Gegründet 2009 mit 7 Mitgliedern in London
- Globales Konsortium getragen von lokalen Institutionen
- Geschäftsstelle an der TIB Hannover
- 2015: 24 Vollmitglieder und 8 assoziierte Mitglieder aus 19 Ländern



DataCite: Ziele

Was?

- Leichterem Zugang zu Forschungsdaten im Internet ermöglichen
- Akzeptanz von Forschungsdaten als relevanter, zitierfähiger Bestandteil des wissenschaftlichen Leistungsausweises stärken
- Datenarchivierung unterstützen, sodass Forschungsergebnisse verifiziert und nachgenutzt werden können

Wie?

- DOI-Registrierung und darauf aufbauende Services
- Entwicklung von Richtlinien und Standards
- Kooperationen und Netzwerke

Richtlinien und Standards

An aerial photograph of the ETH Zurich campus, showing various buildings, a large central dome, and the surrounding city and lake. A semi-transparent white rectangular box is overlaid on the center of the image, containing the title text.

DataCite Metadata Schema



DataCite

DataCite - International Data Citation

DataCite Metadata Schema for the Publication and Citation of Research Data

Version 3.1 June 2015

doi:10.5438/0010

Members of the Metadata Working Group

Joan Starr, California Digital Library (chair of working group)

DataCite Metadata Schema

Table 1: DataCite Mandatory Properties

ID	Property	Obligation
1	Identifier (with type sub-property)	M
2	Creator (with name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M

Table 2: DataCite Recommended and Optional Properties

ID	Property	Obligation
6	Subject (with scheme sub-property)	R
7	Contributor (with type, name identifier, and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
10	ResourceType (with general type description sub-property)	R
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point and box sub-properties)	R



relationType

Description of the relationship of the resource being registered (A) and the related resource (B).

Table 9: Description of relationType

Option	Definition	Example and Usage Notes
IsCitedBy	indicates that B includes A in a citation	Recommended for discovery. <code><relatedIdentifier relatedIdentifierType="DOI"relationType="IsCitedBy">10.4232/10.ASEAS-5.2-1</relatedIdentifier></code>
Cites	indicates that A includes B in a citation	Recommended for discovery. <code><relatedIdentifier relatedIdentifierType="ISBN" relationType="Cites">0761964312</relatedIdentifier></code>
IsSupplementTo	indicates that	Recommended for discovery.

Verlinkung von Artikel und 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.
<http://doi.org/10.1594/PANGAEA.727522>

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, <http://doi.org/10.1016/j.margeo.2004.03.017>

The image shows two overlapping web pages. The top page is the PANGAEA data description for dataset 727522, which links to the article in the bottom page. The bottom page is the ScienceDirect article page for 'Reconstruction of paleoceanography off NW Africa during the last 40,000 years: influence of local and regional factors on sediment accumulation' by Kuhlmann et al. (2004). The PANGAEA page includes a citation, abstract, and project information. The ScienceDirect page includes the article title, authors, DOI, and abstract.

PANGAEA 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, Heide (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 $\delta^{18}O$ and ^{14}C 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 (AR) of the Canary Islands region as an indicator of shifts in the upwelling influenced areas for the Holocene (0-12 ky), the deglaciation (12-16 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 (GeoB) ; Center for Marine Environmental Sciences (MARUM) ;

Coverage: Median Latitude: 28.29307 ° Median Longitude: -13.949662 ° South-bound Latitude: -35.250000 ° West-bound Longitude: -76.000000 ° North-bound Latitude: 32.703336 ° East-bound Longitude: -13.286730

Date/Time Start: 1995-05-28T00:00:00 **Date/Time End:** 1999-10-19T00:00:00

Event(s): **Oe0B3344-2** ° Latitude: -35.250000 ° Longitude: -76.000000 ° Campaign: 1995-05-28T00:00:00 ° Elevation: 4300.0 m ° Location: South-East Pacific ° ; **Oe0B4205-2** ° Latitude: 32.100000 ° Longitude: -11.648333 ° Device: M3711 ° Campaign: M3711 ° Basin: Meteor (1986) ° Device: M3711 ° Campaign: M3711 ° Basin: Meteor (1986) ° ; **Oe0B4206-1** ° Latitude: 31.486333 ° Longitude: -11.015000 ° Device: M3711 ° Campaign: M3711 ° Basin: Meteor (1986) ° ;

License: Creative Commons Attribution 3.0 Unported

ScienceDirect Article:

Marine Geology
Volume 207, Issues 1-4, 18 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

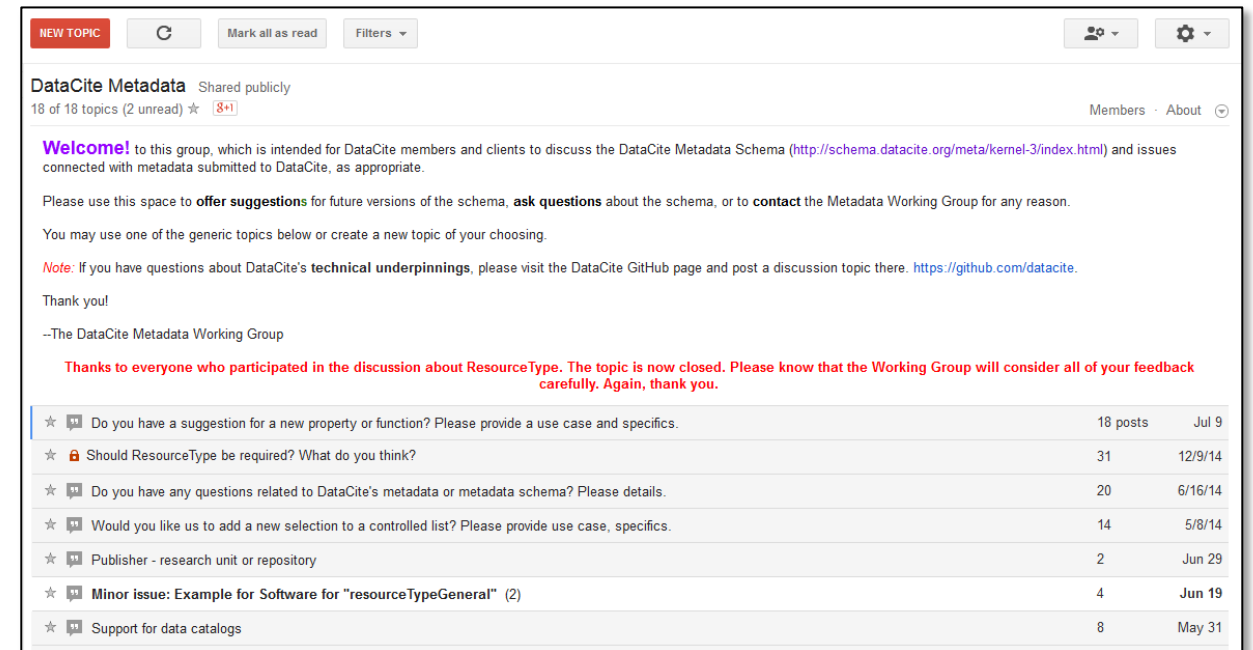
H. Kuhlmann, T. Freudenthal, P. Helmke, H. Meggers

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 $\delta^{18}O$ and ^{14}C 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 (AR) of the Canary Islands region as an indicator of shifts in the upwelling influenced areas for the Holocene (0-12 ky), the deglaciation (12-16 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.

AG Metadaten

- Weiterentwicklung des DataCite Metadaten-Schemas
- ca. 15 Mitglieder
- Monatliche Telefonkonferenzen
- Google Group für Community-Input



NEW TOPIC

DataCite Metadata Shared publicly
18 of 18 topics (2 unread) ★ **8+**

Welcome! to this group, which is intended for DataCite members and clients to discuss the DataCite Metadata Schema (<http://schema.datacite.org/meta/kernel-3/index.html>) and issues connected with metadata submitted to DataCite, as appropriate.

Please use this space to **offer suggestions** for future versions of the schema, **ask questions** about the schema, or to **contact** the Metadata Working Group for any reason.

You may use one of the generic topics below or create a new topic of your choosing.

Note: If you have questions about DataCite's **technical underpinnings**, please visit the DataCite GitHub page and post a discussion topic there. <https://github.com/datacite>.

Thank you!
--The DataCite Metadata Working Group

Thanks to everyone who participated in the discussion about ResourceType. The topic is now closed. Please know that the Working Group will consider all of your feedback carefully. Again, thank you.

★ <input type="button" value="🗨️"/> Do you have a suggestion for a new property or function? Please provide a use case and specifics.	18 posts	Jul 9
★ <input type="button" value="🔒"/> Should ResourceType be required? What do you think?	31	12/9/14
★ <input type="button" value="🗨️"/> Do you have any questions related to DataCite's metadata or metadata schema? Please details.	20	6/16/14
★ <input type="button" value="🗨️"/> Would you like us to add a new selection to a controlled list? Please provide use case, specifics.	14	5/8/14
★ <input type="button" value="🗨️"/> Publisher - research unit or repository	2	Jun 29
★ <input type="button" value="🗨️"/> Minor issue: Example for Software for "resourceTypeGeneral" (2)	4	Jun 19
★ <input type="button" value="🗨️"/> Support for data catalogs	8	May 31

<https://www.datacite.org/working-groups/metadata-working-group.html>

AG Policy and Best Practices

- Entwicklung von Richtlinien und Policies für DataCite Mitglieder und Kunden
- Verbreitung von Best Practices, zum Beispiel zum Zitieren von Forschungsdaten
- Monitoring neuer Entwicklungen im Bereich der Forschungsdatenpublikation
- Unterstützung des DataCite Boards bei der Umsetzung neuer strategischer Vorhaben
- Monatliche Telefonkonferenzen, ca. 8 Mitglieder

<https://www.datacite.org/working-groups/policy-and-best-practices-working-group-pbpwg.html>

Forschungsdaten zitieren

Empfehlung:

Creator (PublicationYear): Title. Version. Publisher. ResourceType.
Identifier

Beispiel:

Swaminathan, R., Ramya, T., Karthik, C.S. (2013): Contortrostatin-
Reprolysin Domain Structure. Swiss Institute of Bioinformatics.
<http://doi.org/10.5452/ma-c12zs>

Siehe auch: [Joint Declaration of Data Citation Principles](#)

Dynamische Datensätze zitieren

- Referenzieren eines bestimmten Abschnitts eines Datensatzes
- Referenzieren eines Snapshots (ganzer Datensatz zu einem bestimmten Zeitpunkt)
- Referenzieren eines kontinuierlich aktualisierten Datensatzes mit Zugriffszeitpunkt (nur möglich, wenn bereits vorhandene Daten nicht verändert werden)

Landing Pages

- enthalten
 - **vollständiges Zitat** inkl. DOI-Name
 - beschreibende **Metadaten**
 - **Zugang** zum eigentlichen Forschungsdatensatz
 - Information über allfällige **Nutzungseinschränkungen**
 - Bereitstellung von **Software** oder **Kontextinformationen**
- lesbar für Menschen und Maschinen!

Kooperationen und Netzwerke

An aerial photograph of the ETH Zurich campus, showing various buildings, a large central dome, and the surrounding city and lake. A semi-transparent white rectangular area is overlaid on the center of the image, containing the title text.

DataCite & ORCID

- **ODIN:** ORCID and DataCite Interoperability Network
 - FP7-Projekt, 2012-2014
- **THOR:** Technical and Human infrastructure for Open Research
 - Horizon2020 Projekt, 2015-2017
- **Ziele:** Interoperabilität verbessern, Software-Prototypen, Integration von Services




DataCite & CrossRef

■ Content Negotiation

- <http://crosscite.org/cn/>
- Maschineller Zugriff auf verschiedene Metadatenformate eines DOI-Objekts

■ Citation Formatter

- <http://crosscite.org/citeproc>
- erzeugt Zitate in über 100 Zitierstilen mittels CrossRef- oder DataCite-DOIs



The screenshot shows the 'DOI Citation Formatter beta' interface. It features the DataCite logo on the left and logos for CrossRef, mEDRA, and Chinese DOI on the right. Below the logos, there is a form with a 'DOI:' label, an empty text input field, a 'Style:' dropdown menu set to 'apa', a 'Locale:' dropdown menu set to 'en-US', and a 'Format' button.

DataCite & Thomson Reuters

- **Data Citation Index:**
 - TR harvested Metadaten von Forschungsdatenrepositorien via DataCite
- **Vorteile für Kunden:**
 - Sichtbarkeit im Web of Science inkl. Verlinkung zu Publikationen
 - Bereitstellung eines zusätzlichen Metadaten-Feed nicht notwendig
- **Vorteile für DataCite:**
 - Indexierung im DCI erhöht Motivation zur Lieferung von qualitativ hochwertigen Metadaten



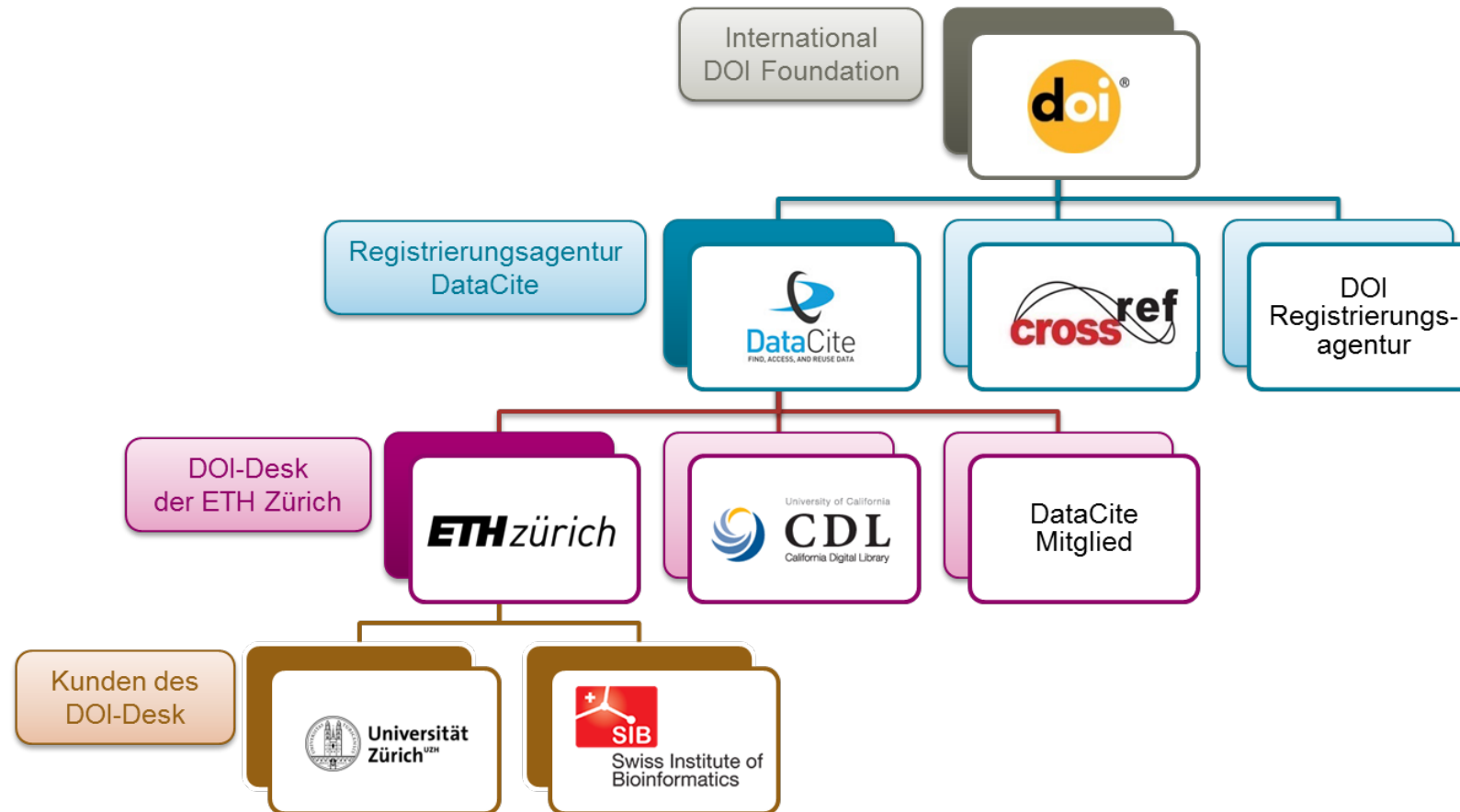
Weitere Kooperationen

- **re3data.org / Databib**
 - ab 2016 Betrieb unter dem Namen **re3data.org** als DataCite-Service
- **Research Data Alliance**
 - Memorandum of Understanding, Organizational Affiliate
- **Force11**
 - Joint Declaration of Data Citation Principles
- **International Association of STM Publishers**
 - Joint Statement to encourage publishers and data centers to link articles and underlying data



Beispiel aus der Praxis

DOI-Desk der ETH Zürich



IT'IS Foundation

The screenshot shows the IT'IS Foundation website. At the top left is the logo 'IT'IS FOUNDATION'. To the right is a search bar with the text 'Search' and a magnifying glass icon. Below the logo is a navigation menu with the following items: NEWS, VIRTUAL POPULATION (highlighted in red), IT'IS FOR HEALTH, EM RESEARCH, SERVICES, and ABOUT. Under 'VIRTUAL POPULATION', there are sub-links: VIRTUAL POPULATION, REGIONAL HUMAN MODELS (highlighted in red), ANIMAL MODELS, and TISSUE PROPERTIES. The main content area features a large banner for 'REGIONAL HUMAN MODELS' with a 3D wireframe model of a human torso. Below the banner is a sub-navigation menu: OVERVIEW, MIDA MODEL (highlighted in red), and TRACEABILITY. The main content is titled 'MIDA Model V1.0' and includes a 3D model of a human head with internal structures. To the right of the model is a list of metadata: DOI: 10.13099/ViP-MIDA-V1.0, Creator: FDA & IT'IS Foundation, Title: MIDA Model V1.0, Publisher: IT'IS Foundation, Release Date: 22.04.2015, and Inquiries: US Food and Drug Administration or Virtual Population Group. Below the model is a short description: 'The MIDA model: one of the the most detailed image-based anatomical head models available for computational life sciences'. To the right of the main content is a sidebar with a 'MIDA V1.0' header, an 'INQUIRIES' section with links to 'US Food and Drug Administration' and 'Virtual Population Group', and a 'COLLABORATORS' section with logos for 'FDA' and 'ETH'. At the bottom right of the main content area, there is a label 'Affiliation*'.

Danke für Ihre Aufmerksamkeit!