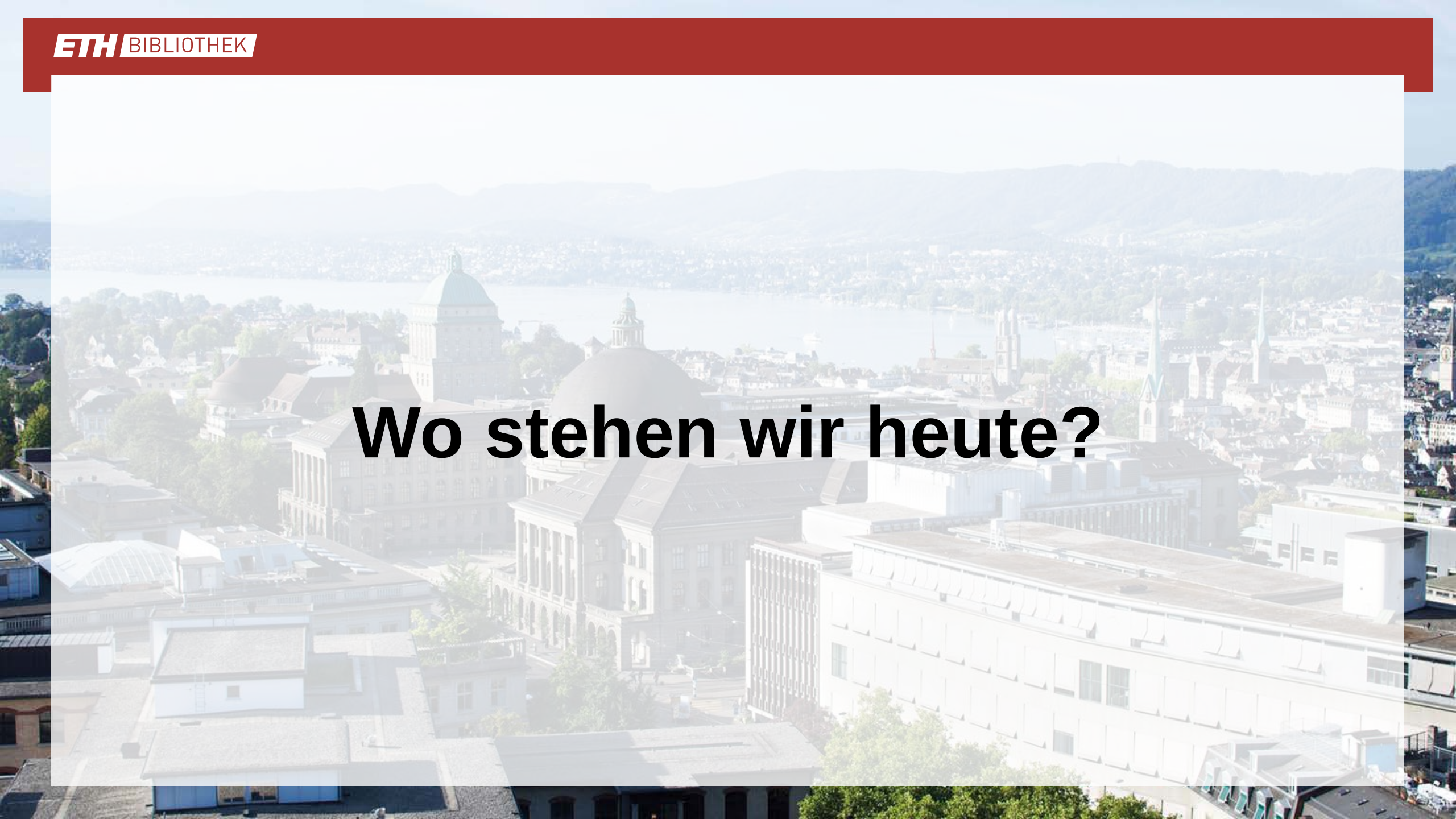
ETH BIBLIOTHEK

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



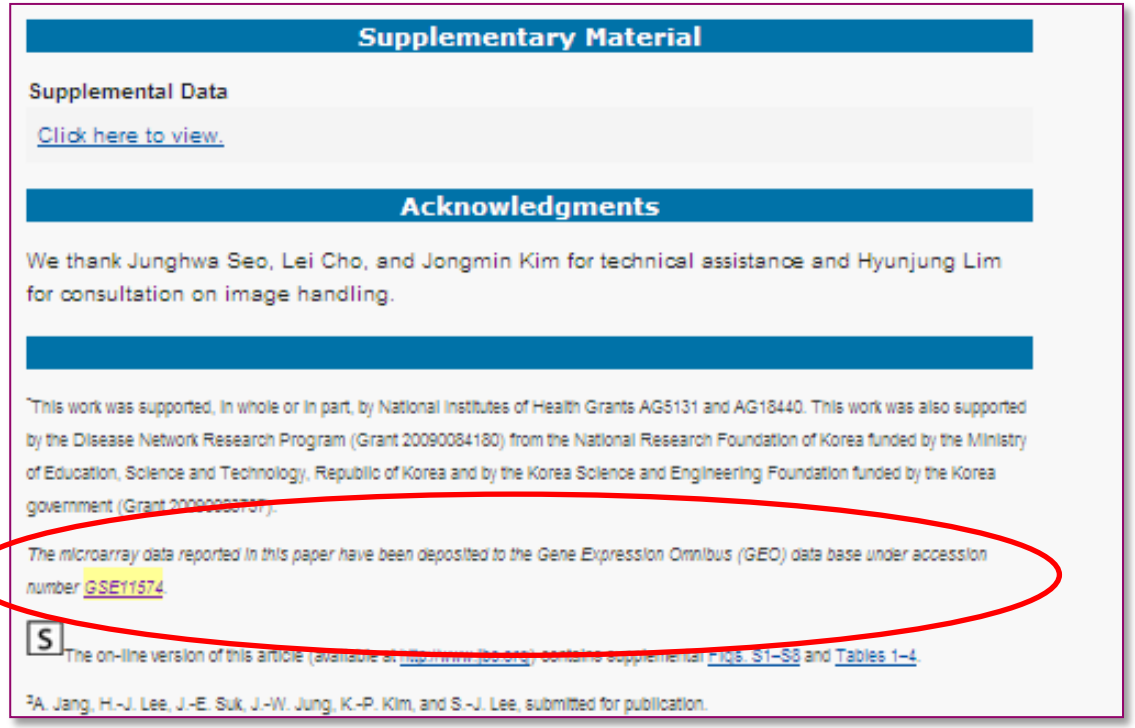
Wo stehen wir heute?

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 [illegible]).

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 [illegible]) contains supplemental Figs. S1–S8 and Tables 1–4.

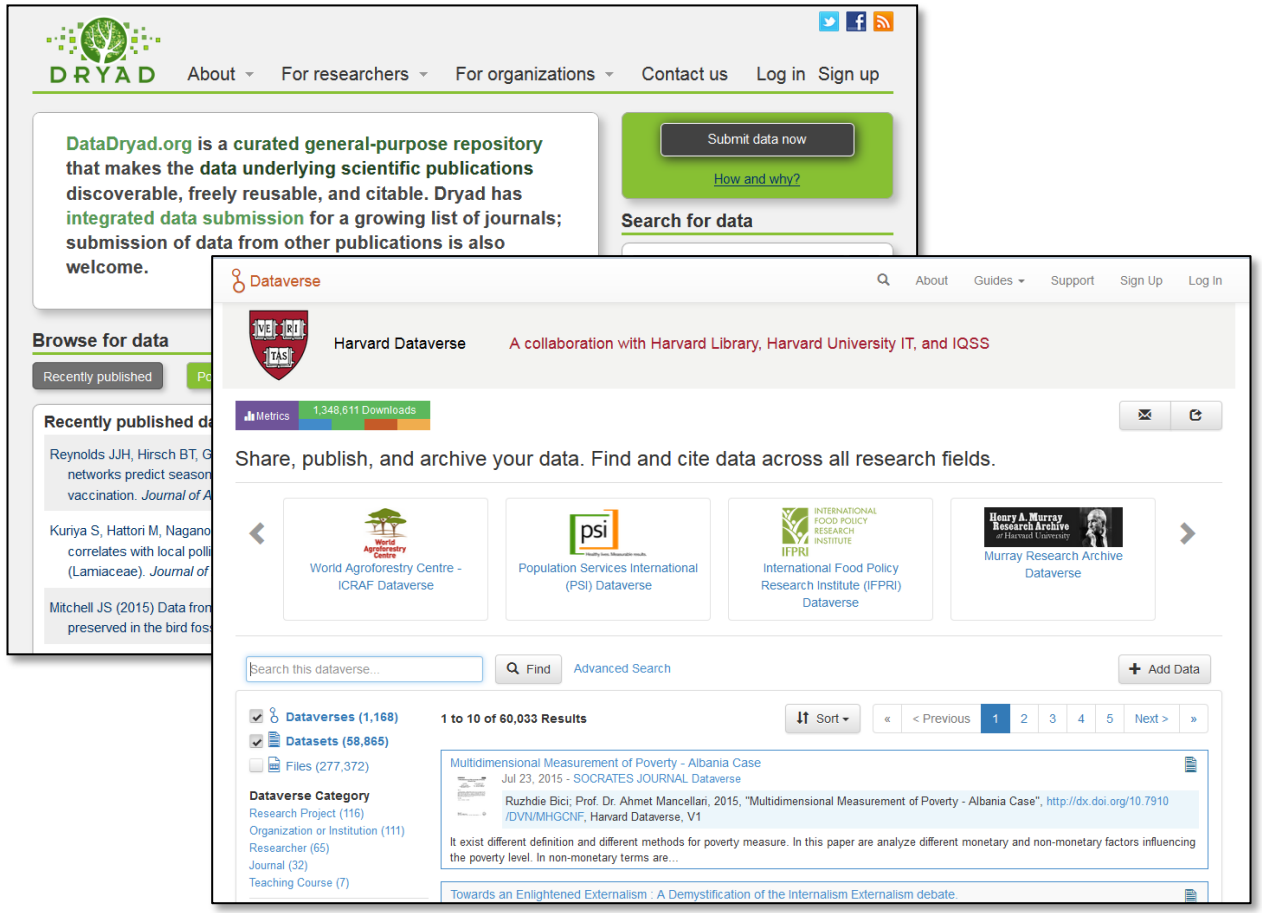
[3]A. Jang, H.-J. Lee, J.-E. Suk, J.-W. Jung, K.-P. Kim, and S.-J. Lee, submitted for publication.

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

Wo möchten wir hin?

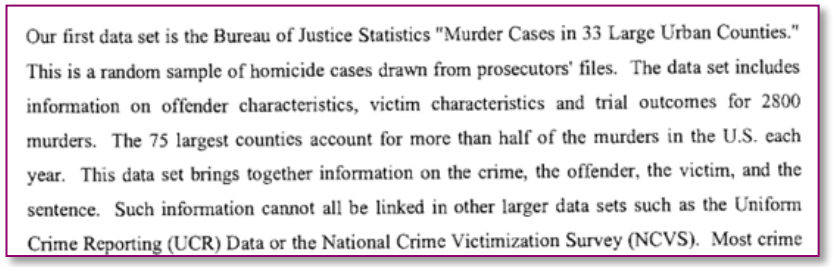
Publikation in vertrauenswürdigen Repositorien



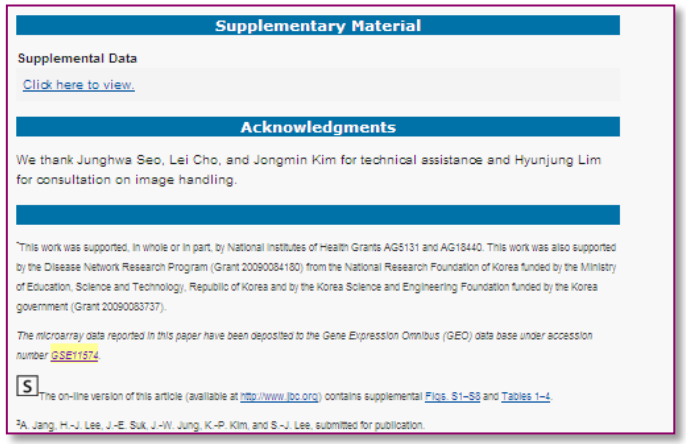
- Vertrauenswürdiger Betreiber
- Sicherstellung des langfristigen Zugriffs
- Bereitstellung von Kontextinformationen (Metadaten)
- Persistente Identifizierung der Datensätze (z.B. via DOI)

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



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.

[illegible]

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.

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

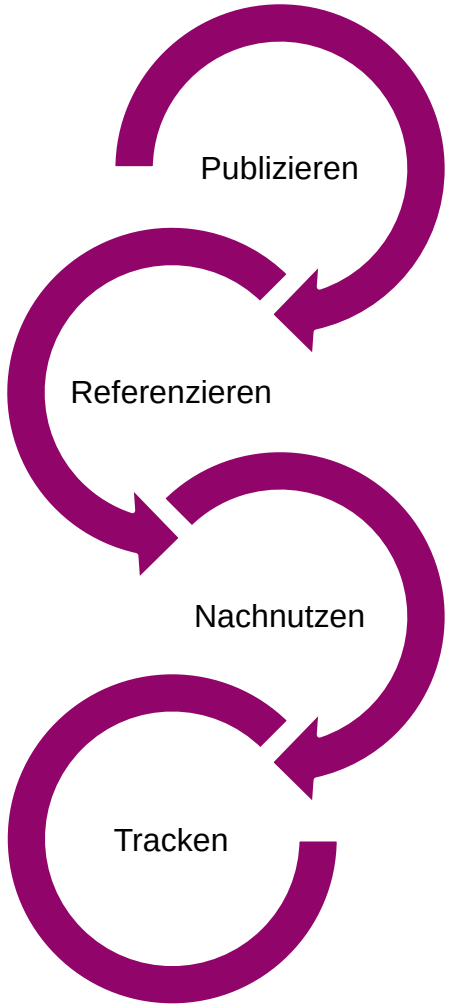
[illegible]A. Jang, H.-J. Lee, J.-E. Suk, J.-W. Jung, K.-P. Kim, and S.-J. Lee, submitted for publication.

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

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?




ETH BIBLIOTHEK
Wie kann DataCite dazu beitragen?

Das DOI-System

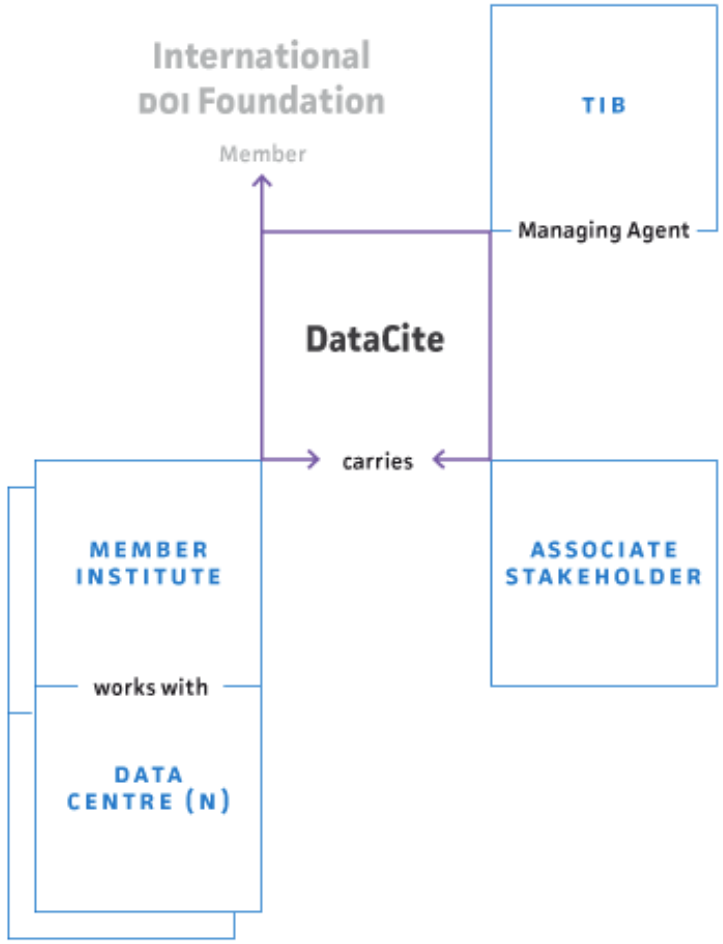
- **International DOI Foundation** (IDF)
  - 1998 gegründet
  - besteht derzeit aus 9 Registrierungsagenturen, darunter CrossRef und DataCite

- **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?

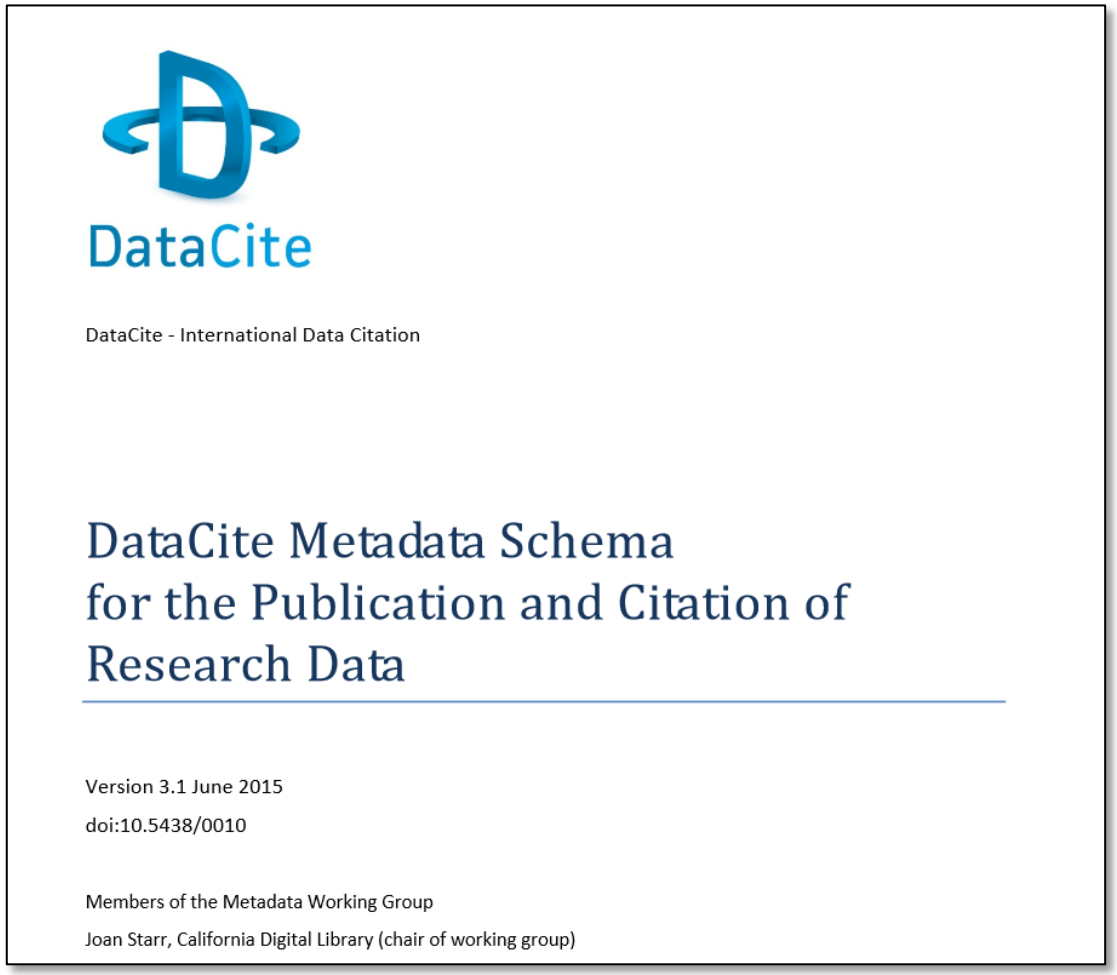
- Leichteren Zugang zu Forschungsdaten im Internet ermöglichen
- Akzeptanz von Forschungsdaten als relevanter, zitierfähiger Bestandteil des wissenschaftlichen Leistungsausweises stärken
- Datenarchivierung unterstützen, sodass Forschungsresultate verifiziert und nachgenutzt werden können

Wie?

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


ETH BIBLIOTHEK
Richtlinien und Standards

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

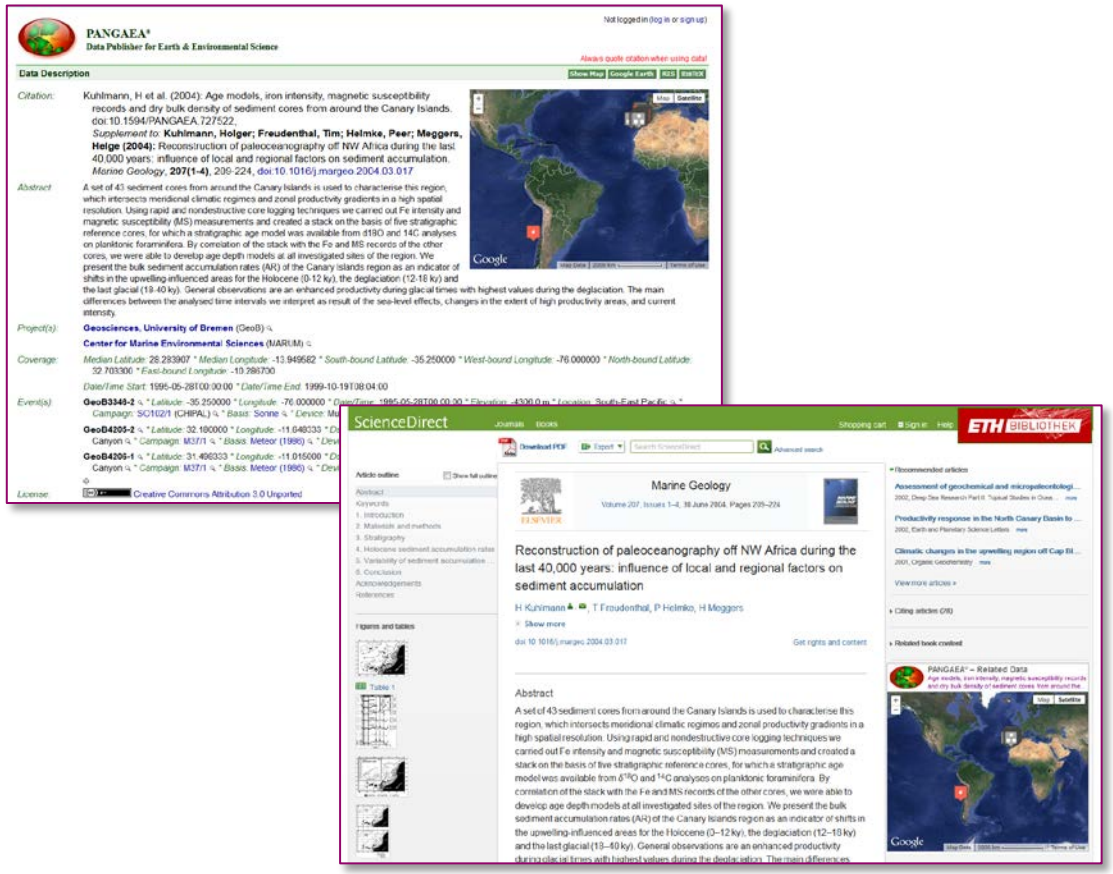
Verlinkung von Artikel und Forschungsdaten

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.org10.1594/PANGAEA.727522

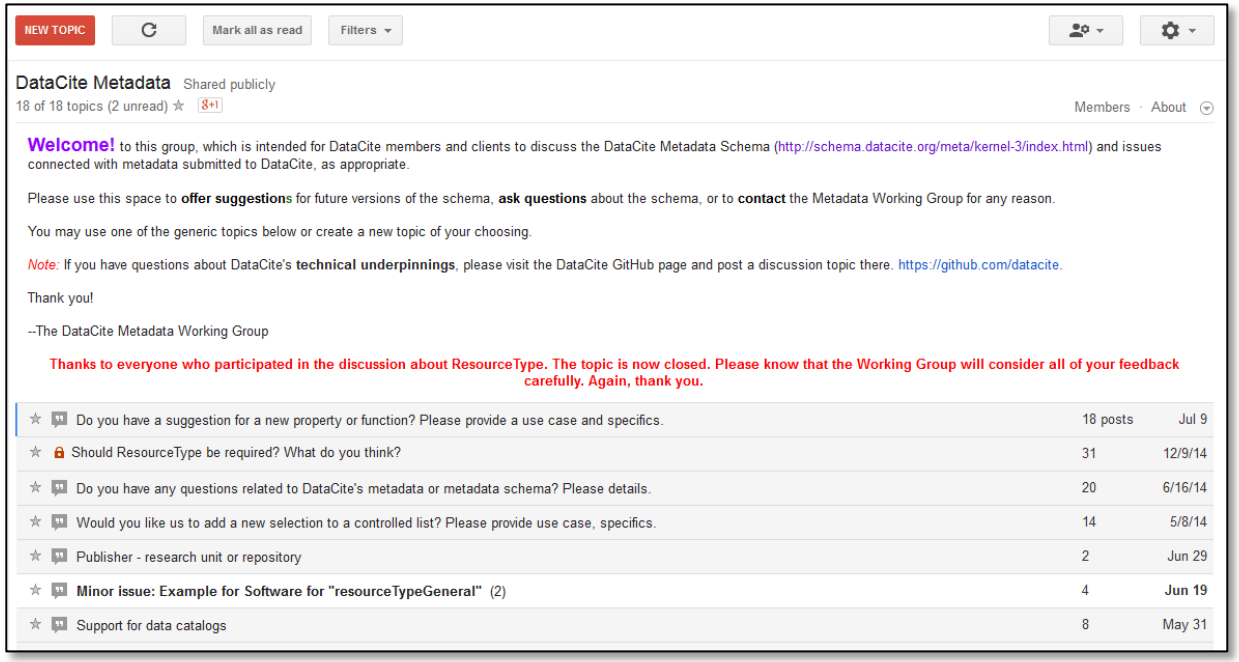
in folgendem Artikel analysiert

- 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



AG Metadaten

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



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

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

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

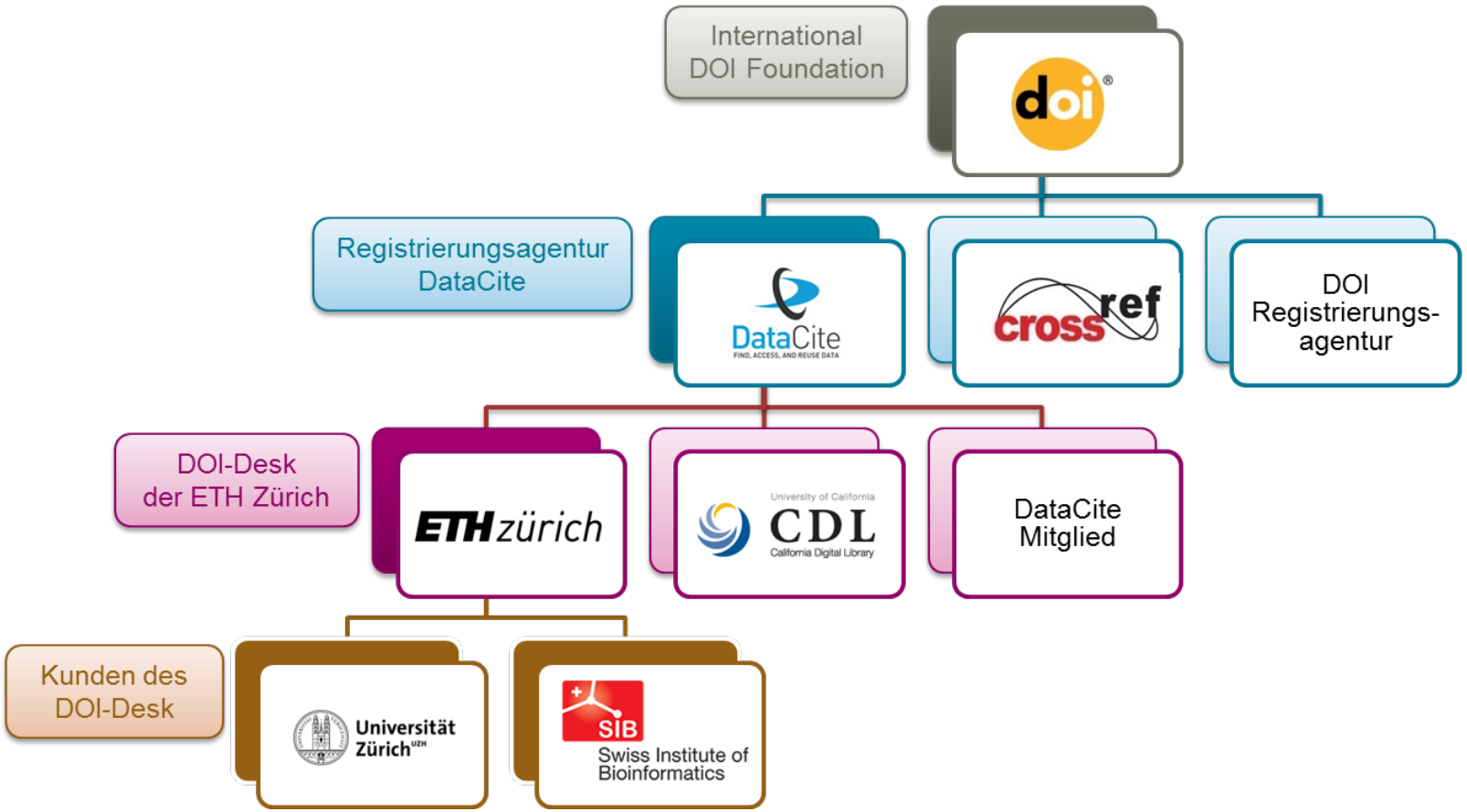
- **Vorteile für DataCite:**
  - Indexierung im DCI erhöht Motivation zur Lieferung von qualitativ hochwertigen Metadaten

Weitere Kooperationen

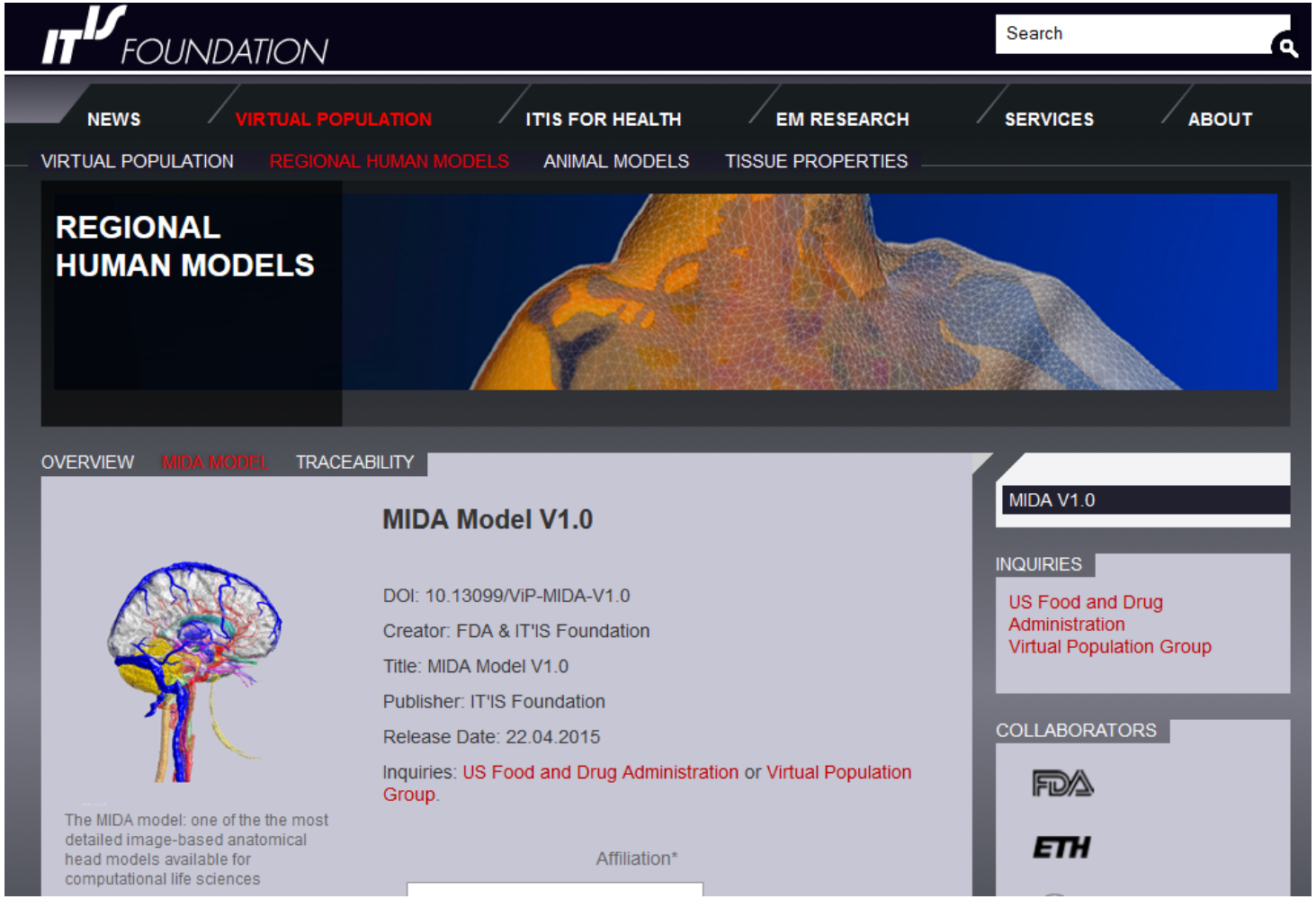
- **re3data.org / Databib**
  - ab 2016 Betrieb unter dem Namen **re3data.org** als DataCite-Service
- **Research Data Alliance**
  - Memorandum of Understanding, Organizational Affiliate
- **Force11**
  - Joint Declaration of Data Citation Principles
- **International Association of STM Publishers**
  - 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



Danke für Ihre Aufmerksamkeit!