RESEARCH DATA ALLIANCE

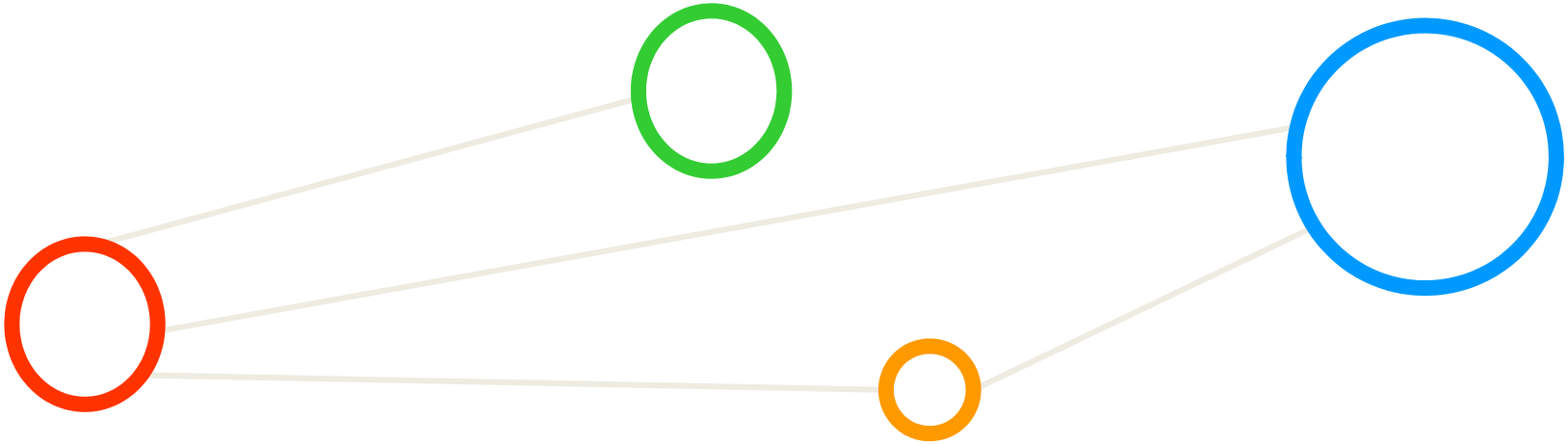
Eine Weltweite Initiative

zur Steigerung von Interoperabilität und Wiederverwendbarkeit von Forschungsdaten

Prof. Dr. – Ing. Morris Riedel, Helmholtz Association
Co-Chair RDA Big Data Interest Group

research data sharing without barriers
rd-alliance.org

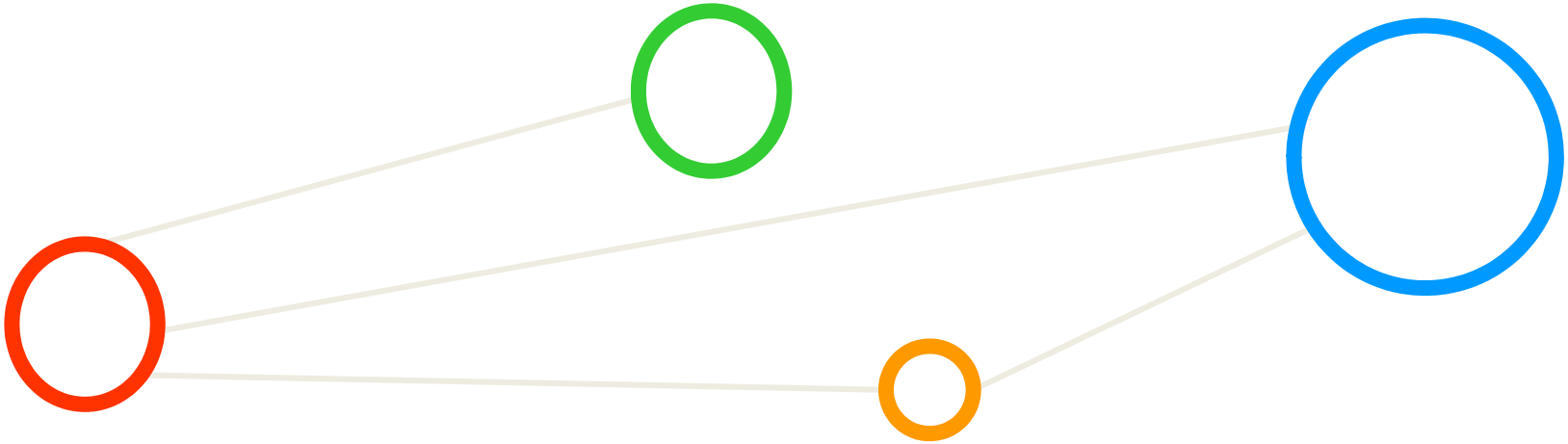
Gliederung



Gliederung

- Überblick der Initiative
- Arbeitsweise
- Erste Resultate
- Umsetzungsbeispiele
  - Impuls zu unserer Diskussion: „Archivierung von Forschungsdaten“
- Einladung zur Mitarbeit
- Zusammenfassung
- Referenzen

Überblick der Initiative



Grundlegende Ideen der Initiative

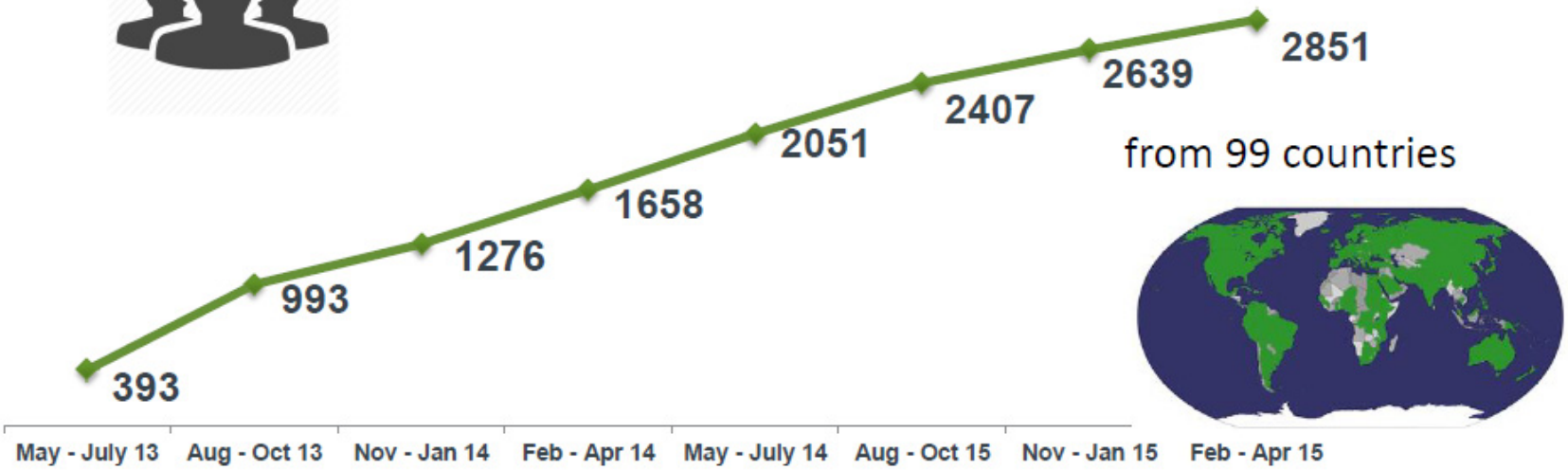
[1] RDA Website (Online)

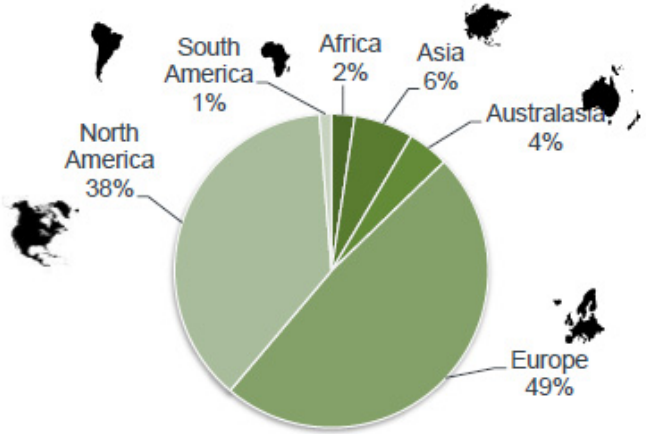
- „RDA Vision“ verwirklichen
  - Weltweit: Überwindung von Ländergrenzen
  - Forscher können frei verfügbare Daten austauschen
  - Austausch über Technologiegrenzen und Disziplinen
- „RDA Mission“
  - Aufbau von „sozialen und technischen Brücken“ für den Datenaustausch
- Angestrebte Resultate
  - Erstellen von Empfehlungen für Organisationen aus diversen Bereichen
  - Beispiel: Klärung und gemeinsames Verständis von Terminologie
  - Beispiel: Vorgehensweisen und gemeinsame Ansätze besser verstehen

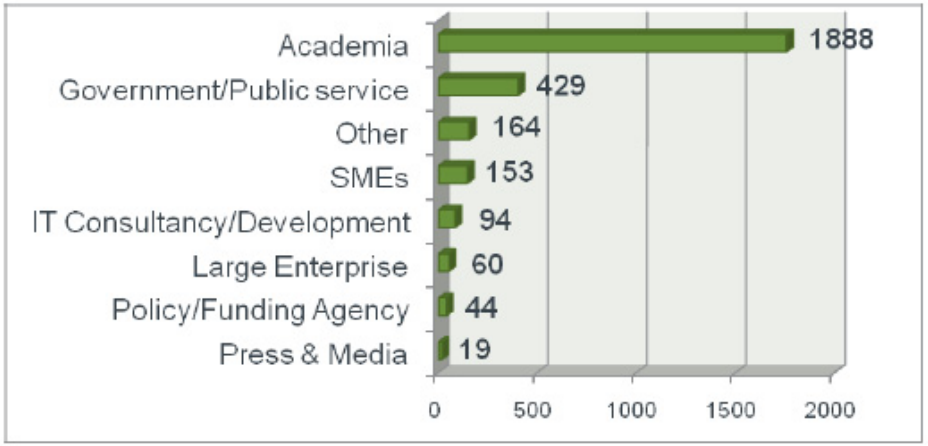
Wichtig: RDA fördert zwar den Datenaustausch – sammelt aber selbst keine Daten ein!

Internationale Mitglieder

Totale Anzahl Mitglieder heute: ~3000







Mitglieder Organisationen

in 2016

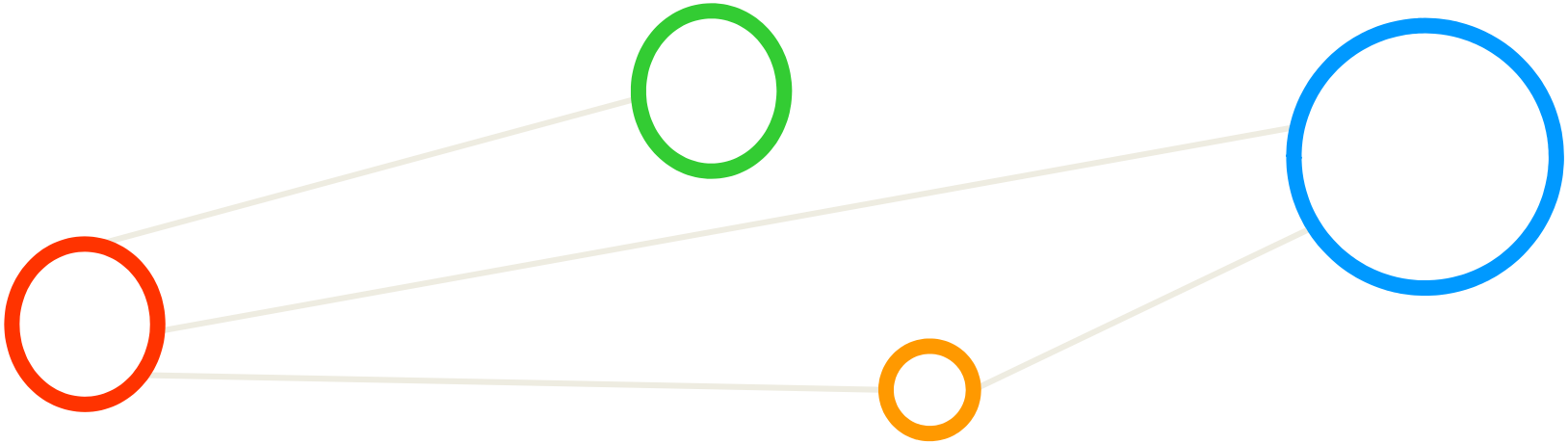
Totale Anzahl Organisationen: ~38

RDA Deutschland

- Idee: Kräfte in Deutschland bündeln
  - Diskussionen zur langfristigen RDA Teilnahme (auch Gelder)
  - Kleinere Tagungen in Deutschland
  - Trainings anbieten zur Umsetzung!
  - Repräsentation größerer Forschungsgemeinschaften stärken
  - Bspw. Helmholtz Gemeinschaft
  - Bspw. Max- Planck Gesellschaft

- Zwei Tagungen
  1. 2014-11 – Potsdam
  2. 2015-11 – Potsdam

Arbeitsweise



Zwei Gruppenarten

- Interessen – Gruppen (engl. ‘interest groups‘)
  - Ziel: Identifikation gemeinsamer Schwerpunktthemen
  - Zeitlich unbefristet, oft aus ‘Communities‘ heraus
  - Keine direkten Resultate angestrebt

Totale Anzahl Interessen - Gruppen: ~45

Im Idealfall ergeben sich 1-3 konkrete Arbeitsgruppen

- Arbeits – Gruppen (engl. ‘working groups‘)
  - Ziel: Konkretes Resultat in einem Schwerpunktthema
  - Zeitlich befristet, oft fokussiert auf ein Problem („case statements“)
  - Klar meßbare Resultate verlangt

Totale Anzahl Arbeits - Gruppen: ~27

Gruppenbeispiele

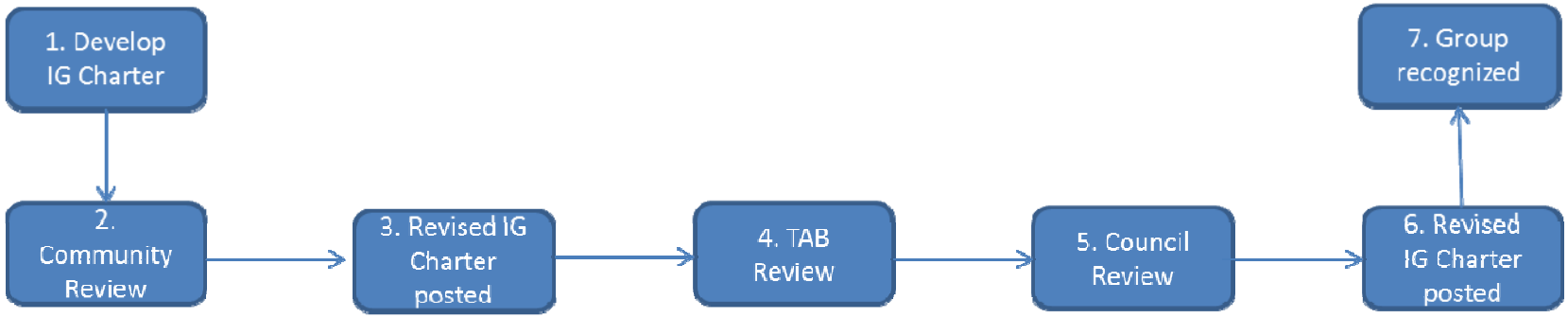
- Reproducability IG
- Active Data Management Plans IG
- Big Data IG
- Biodiversity Data Integration IG
- Chemistry Research Data IG
- Data Fabric IG
- Data for Development IG
- Data Rescue IG
- …
- Domain Repositories IG
- Ethics and Social Aspects of Data IG
- Federated Identity Management IG
- Global Water Information IG

- Agrisemantics WG
- Array Database WG
- BioSharingRegistry WG
- Brokering Framework WG
- Brokering Governance WG
- Data Citation WG
- Data Description Registry Interoperability WG
- …
- Data Foundation and Terminology WG
- Data Type Registries WG
- Data Security and Trust WG
- Wheat Data Interoperability WG

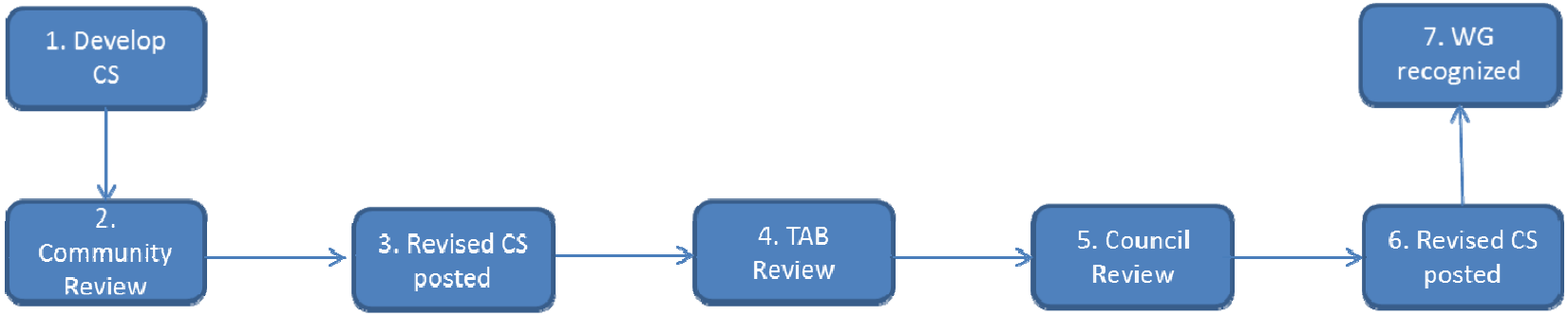
Erstellung neuer Gruppen

Gewähltes Technical Advisory Board (TAB) entscheidet zusamment mit RDA Council

- Prozess für Interessen – Gruppen



- Prozess für Arbeits – Gruppen

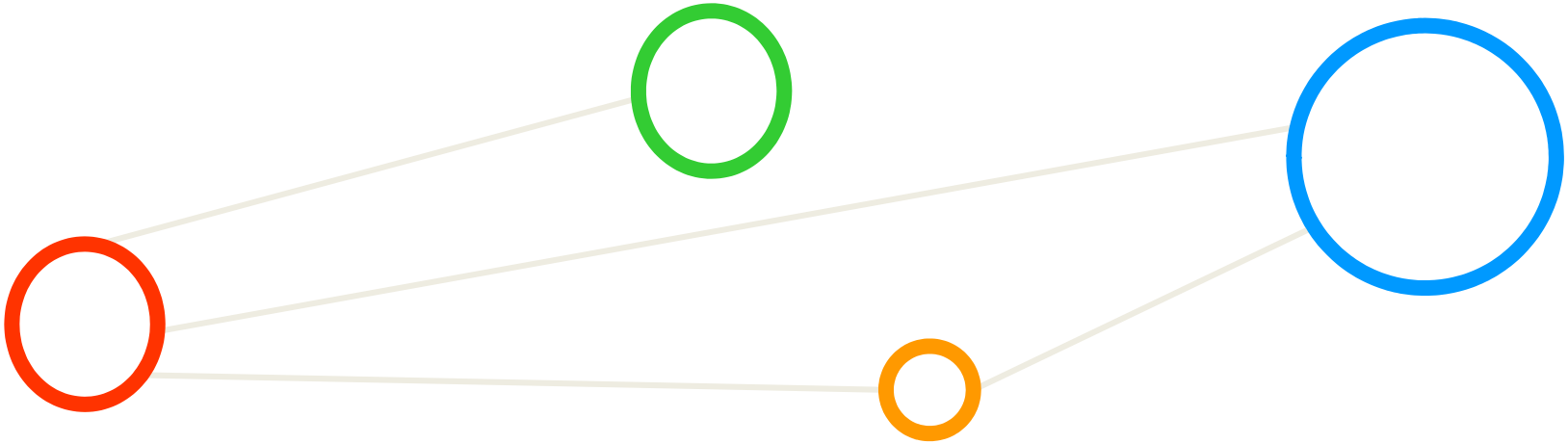


Tagungen

- Ziele
  - Gruppentreffen mit schnellerem Fortschritt, Sitzungen werden beantragt
  - Bündelung neuer Ideen zu Gruppen, bspw. „Birds of Feathers (BOFs)“
  - Zeigen von Resultaten und Umsetzungen – auch „Special Focus“ / Events

- Bisherige Tagungen
  1. 03-2013 – Göteborg, Schweden
  2. 09-2013 – Washington, USA
  3. 03-2014 – Dublin, Irland
  4. 09-2014 – Amsterdam, Niederlande
  5. 03-2015 – San Diego, USA
  6. 09-2015 – Paris, Frankreich
  7. 03-2016 – Tokio, Japan

Erste Resultate



Überblick Erste Resultate

- Einige Arbeitsgruppen sind „fertig“
- Resultate sind in einem „Outputs Booklet“ zusammengefasst: hier nur einige Beispiele
- Alle Resultate sind offen
- Umsetzung in verschiedenen Einrichtungen beginnt
- Auch Interessen – Gruppen haben initiale Resultate
  - Bspw. Big Data IG
  - Idee: www.big-data.tips
    Webseite mit Empfehlungen auch für „nicht Experten“

BIG DATA TOOLS

Using the Hadoop Big Data Tool

BY WWW.BIG-DATA.TIPS · MARCH 11, 2016

Like Share

The Hadoop Big Data tool provides a solution to perform data analysis and analytics on large quantities of data. This is possible, because Hadoop uses an approach of parallelization. In other words the data is not analyzed only on one machine or Laptop, but in parallel at the same time on a large server machine.

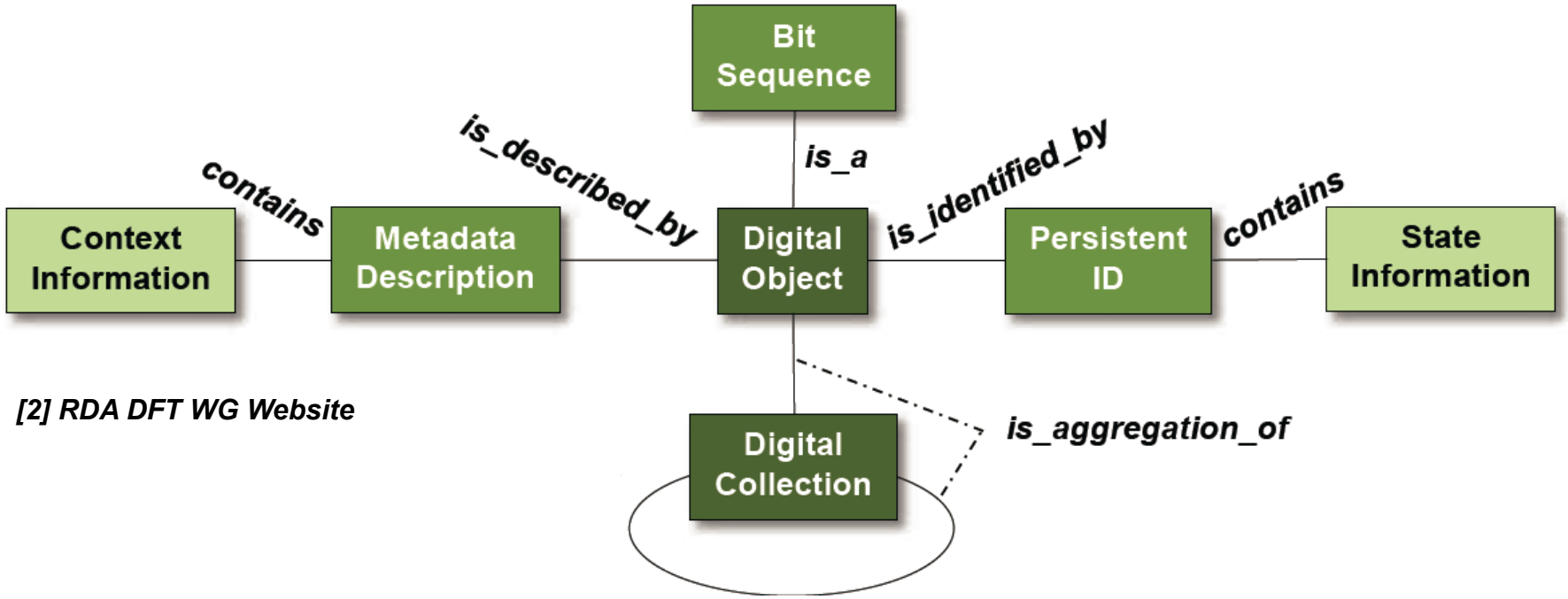
Hadoop is very powerful but it is important to understand that it requires an underlying computing infrastructure that is different from a usual desktop computer or simple laptop. It is thus not a solution like R or Matlab that you simply run on your own computer and therefore users of this tool usually work together with some data center that offers a computing infrastructure for it.

Hadoop Tool Tutorials

There are many tutorials online, but this tutorial offers also a good Background

Erste Resultate – Einfaches Datenmodel

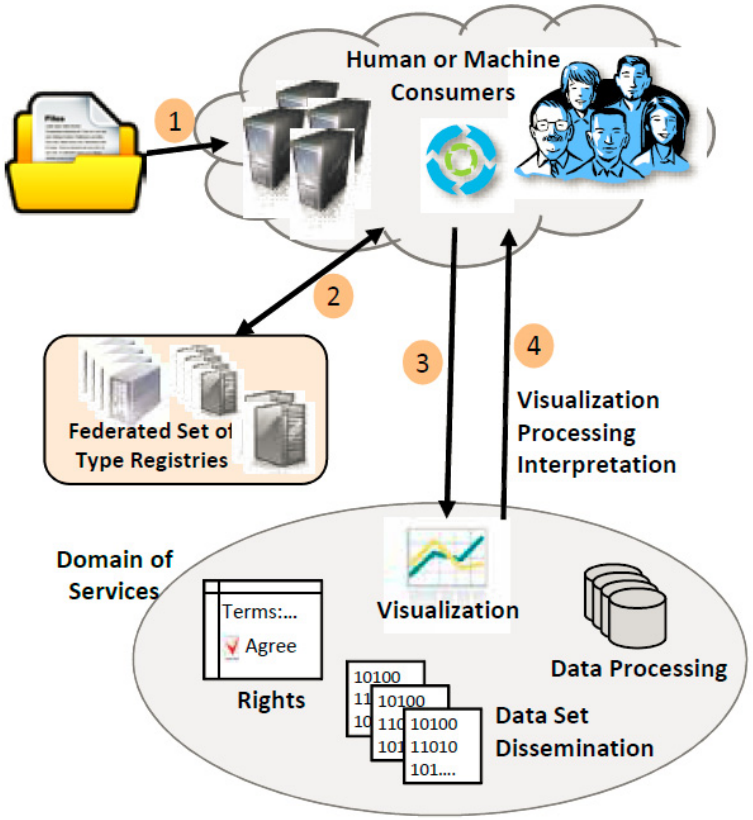
- Data Foundation and Terminology WG Resultat
  - Ziel: Einigung auf ein sehr einfaches Datenmodell wäre prima in der Praxis
  - Diskussion der Frage: „Was ist ein digitales Objekt nun genau?“
  - Enthält Definitionen: Persistent Identifier (PID), Digital Object, Metadata, …



[2] RDA DFT WG Website

Erste Resultate – Wissenschaftliche Datentypen

- Data Type Registries WG Resultat
  - Ziel: Einigung auf einen Ansatz damit unbekannte Datentypen automatisch bekannte Datentypen werden (ohne Wissen des Nutzers)
  - Diskussion der Frage: „Wie kann ich einen Datentyp nun genau festlegen?“
  - Enthält Ansatz wie „Type Registries“ und Datentypen Service Vorlage zur Implementierung (ähnlich wie MIME Typen im Browser)
  - Prototyp der Implementierung online: http://typeregistry.org/



[3] RDA DTR WG Website

Erste Resultate – Wissenschaftliche Datentypen

http://typeregistry.org/registrar/ Data Type Registry

Data Type Registry About Scope API Primitives Derived All

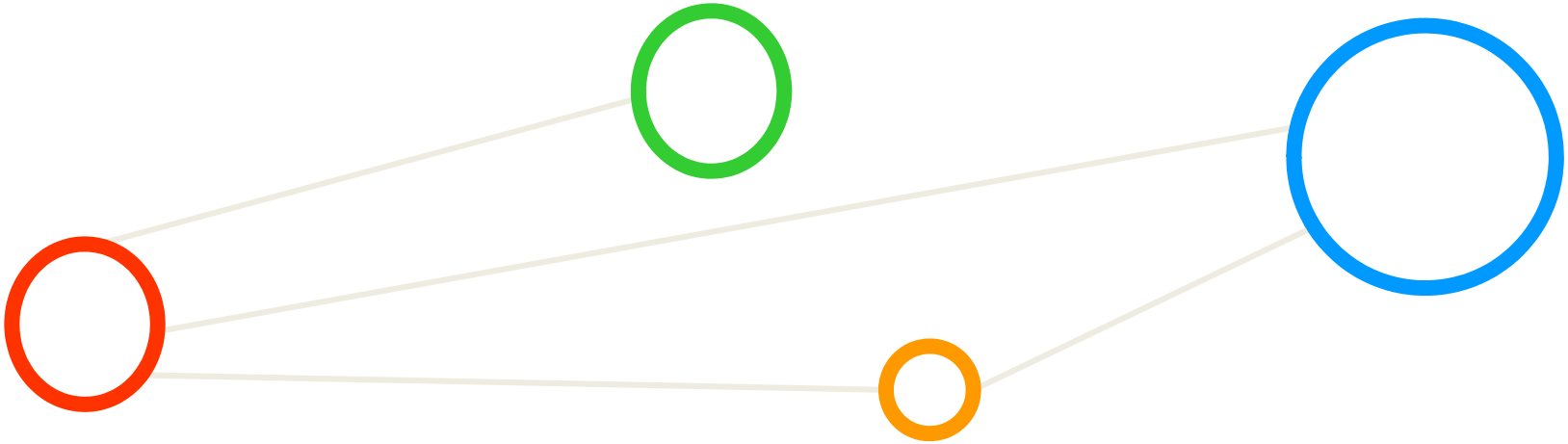
Search Search Create

Retrieve JSON ✖

- Boolean
  Id: 11314.3/b76ff4906f33c2dd97dd
  A computer representation of a value representing truth values in logic: True or False; Yes or No; and 1 or 0.
- XML Schema Repository XSD
  Id: 11314.3/xsd-record
  This is a schema for use in a repository for XSD files.
- MDCS xray-siegbahn-type
  Id: 11314.3/d422e23d8dcfcd1a9e995b0bdca63654
  Siegbahn notation is used in X-ray spectroscopy to name the spectral lines. This is a XML schema representation for XML serialization. Keywords: K-alpha; Kalpha; K-beta; Kbeta; x-ray source
- MDCS physical-quantity-type
  Id: 11314.3/f3b9d004f79648c20992d6f4d0a72256
- Miller Index Direction Type
  Id: 11314.3/cd9660baaab138b9661a
  This reusable type may be used to specify a crystallographic direction or family of equivalent directions in a crystal Bravais lattice. As all enclosed elements are optional, this type may be used for two-dimensional representations (h, k) or three-dimensional representations in either the three-index (h, k, l) or four index (h, k, i, l) notation.
- Miller Index Plane Type
  Id: 11314.3/782c96247904167703cb
  This reusable type may be used to specify a crystallographic plane or family of equivalent planes in a crystal Bravais lattice. As all enclosed elements are optional, this type may be used for two-dimensional representations (h, k) or three-dimensional representations in either the three-index (h, k, l) or four index (h, k, i, l) notation.
- Lidar Point Classification
  Id: 11314.3/70ff9bf267d79998024e
  A type of physical object that has reflected the laser pulse emitted by a sensor during a light detection and ranging (lidar) optical remote-sensing operation.
- MDCS uncertainty-type
  Id: 11314.3/81fad9021bab92efb9bd9fcf684fd77f
- MDCS physical-quantity-L--type
  Id: 11314.3/7f75a53749c6a8b8369b5581729a1724
  Length

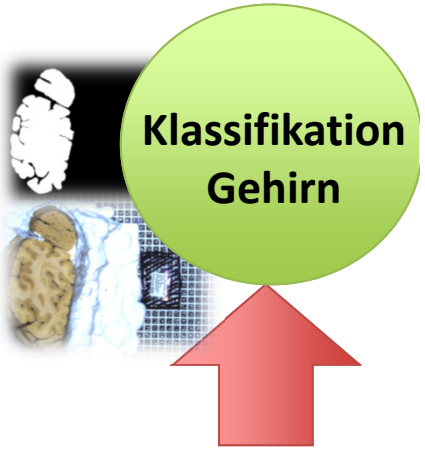
[3] RDA DTR WG Website

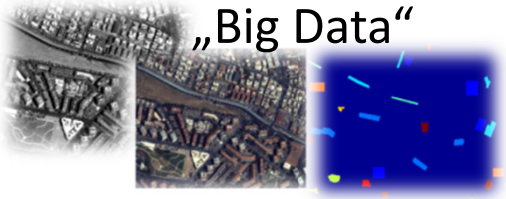
Umsetzungsbeispiele



Umsetzung PID: Archivierung von Forschungsdaten

- Beispiel: Big Data IG Bereich „Big Data Analytics“
  - Bedeutet: Paralleles & Skalierbares Maschinelles Lernen
  - Wiss. Problem: Klassifikation der Erdbodenbedeckung





„Big Data“

Satellitenaufnahmen

[4] G. Cavallaro & M. Riedel et al., ‘On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machines ’, IEEE Journal of Selected Topics in Applied Earth Observation, 2015

Parallel Support Vector Machines (SVM)

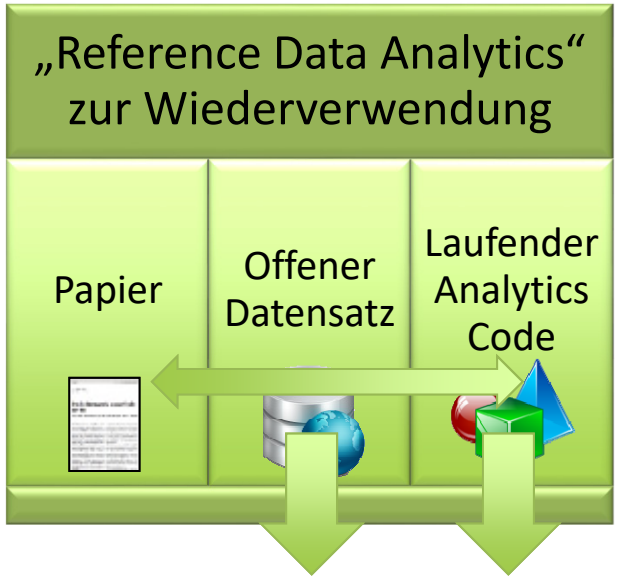
HPC/MPI, Map-Reduce & GPGPUs

Klassifikation Erdboden-Bedeckung

„Peer-Review“

„Best Practices“

Community-basierte Empfehlungen



B2SHARE Store and Share Research Data

πSVM

[5] EUDAT B2SHARE

[6] piSVM

Umsetzung PID: Archivierung von Forschungsdaten

piSVM1.2.1 Analytics Training Indian Pines Images Processed 30 Features 52 Classes

Morris Riedel

;

23 February 2015

https://b2share.eudat.eu

B2SHARE
Store and Share Research Data

[5] EUDAT B2SHARE

Abstract: piSVM version 1.2.1
Deployment JUDGE cluster of the Juelich Supercomputing Centre in Germany
Parallel Support Vector Machine (SVM) classification runs to analyse
remote sensing hyperspectral images
(52 land cover classes, 30 features).
Indian processed:1417x614x30 (training 10% and test)
Data preprocessed with PCA, ESDAP, NWFE
Supplemental material for paper study.
Correspondending dataset (processed) available at:
http://hdl.handle.net/11304/7e8eec8e-ad61-11e4-ac7e-860aa0063d1f

Metadata

Daten und Skriple archiviert und vom Papier <u>per PID referenziert</u> → entspricht Definition <u>RDA „Digital Object“</u> für die Papierdaten

Keyword(s): SVM ; remote sensing ; analytics ; MPI ; classification

The record appears in these collections:
Generic

Data (Bit sequence)

Files

Name	Date	Size	
2226457-indian_processed_training.el.model.model	24 Feb 2015	20.1 MB	Download
Pred-rec172-1-2-1.02236897.02236897	24 Feb 2015	221 Bytes	Download
2236893-checkjob.el	24 Feb 2015	1.2 kB	Download
2234622-results-1-8-4.txt	24 Feb 2015	793.1 kB	Download
2234612-submit-train-record172.sh	24 Feb	600	Download

Export

Export as BibTeX, MARC, MARCXML, DC, EndNote, NLM, RefWorks

PID handle

Metadata

PID:	http://hdl.handle.net/11304/fe145db0-bc5f-11e4-ac7e-860aa0063d1f
Publication:	https://b2share.eudat.eu
Publication Date:	2015-02-23
Licence:	Unspecified

Umsetzung PID: Archivierung von Forschungsdaten

[62] F. Garcia-Vilchez *et al.*, "On the impact of lossy compression on hyper-spectral image classification and unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 253–257, Mar. 2011.

[63] G. Cavallaro and M. Riedel. (2015). *Indian Pines Dataset Raw and Preprocessed* [Online]. Available: http://hdl.handle.net/11304/7e8eec8e-ad61-11e4-ac7e-860aa0063d1f

[64] G. Cavallaro and M. Riedel. (2015). *piSVM Analytics 10-Fold Cross-Validation Results Scenario 1 (Raw)* [Online]. Available: http://hdl.handle.net/11304/163ba8e8-fe60-11e4-8a18-f31aa6f4d448

[65] G. Cavallaro and M. Riedel. (2015). *piSVM analytics 10-fold cross-validation results scenario 2 (processed)* [Online]. Available: http://hdl.handle.net/11304/5bba8e36-fe63-11e4-8a18-f31aa6f4d448

[66] G. Cavallaro and M. Riedel. (2015). *piSVM Analytics Results Scenario 1 (Raw)* [Online]. Available: http://hdl.handle.net/11304/c06a8c7e-fe6c-11e4-8a18-f31aa6f4d448

[67] G. Cavallaro and M. Riedel. (2015). *piSVM Analytics Results Scenario 2 (Processed)* [Online]. Available: http://hdl.handle.net/11304/c528998e-ff7c-11e4-8a18-f31aa6f4d448

handle PID links

PID handle enables persistent link for changing locations (challenge for archiving)

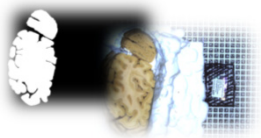
B2SHARE
Store and Share Research Data

[5] EUDAT B2SHARE

more reputation for authors

enables data-sharing

enables reproducability



This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

CAVALLARO *et al.*: ON UNDERSTANDING BIG DATA IMPACTS IN REMOTELY SENSED IMAGE CLASSIFICATION 13

[58] L. Moisan, J. Froment, and J. M. Morel, (1993, Nov.). *Megawave2* [Online]. Available: http://megawave.cmla.ens-cachan.fr/index.php

[59] P. Monasse and F. Guichard, "Fast computation of a contrast-invariant image representation," *IEEE Trans. Image Process.*, vol. 9, no. 5, pp. 860–872, May 2000.

[60] B.-C. Kuo and K.-Y. Chang, "Feature extractions for small sample size classification problem," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 756–764, Mar. 2007.

[61] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction of hyperspectral images," in *Proc. 6th Workshop Hyperspectral Image Signal Process. Evol. Remote Sens. (WHISPERS)*, 2014.

[62] F. Garcia-Vilchez *et al.*, "On the impact of lossy compression on hyper-spectral image classification and unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 253–257, Mar. 2011.

[63] G. Cavallaro and M. Riedel. (2015). *Indian Pines Dataset Raw and Preprocessed* [Online]. Available: http://hdl.handle.net/11304/7e8eec8e-ad61-11e4-ac7e-860aa0063d1f

[64] G. Cavallaro and M. Riedel. (2015). *piSVM Analytics 10-Fold Cross-Validation Results Scenario 1 (Raw)* [Online]. Available: http://hdl.handle.net/11304/163ba8e8-fe60-11e4-8a18-f31aa6f4d448

[65] G. Cavallaro and M. Riedel. (2015). *piSVM analytics 10-fold cross-validation results scenario 2 (processed)* [Online]. Available: http://hdl.handle.net/11304/5bba8e36-fe63-11e4-8a18-f31aa6f4d448

[66] G. Cavallaro and M. Riedel. (2015). *piSVM Analytics Results Scenario 1 (Raw)* [Online]. Available: http://hdl.handle.net/11304/c06a8c7e-fe6c-11e4-8a18-f31aa6f4d448

[67] G. Cavallaro and M. Riedel. (2015). *piSVM Analytics Results Scenario 2 (Processed)* [Online]. Available: http://hdl.handle.net/11304/c528998e-ff7c-11e4-8a18-f31aa6f4d448

Gabriele Cavallaro (S'14) received the B.S. and M.S. degrees in telecommunication engineering from the University of Trento, Trento, Italy, in 2011 and 2013, respectively. Currently, he is pursuing the Ph.D. degree with the University of Iceland, Reykjavik, Iceland.

His research interests include remote sensing and analysis of very high geometrical and spectral resolution images with the current focus on mathematical morphology and high-performance computing.

Morris Riedel (M'10) received the Ph.D. degree from Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, and started the work in parallel and distributed systems in the field of scientific visualization and c… applications on larg…

He is an Adj… the School of En… University of Icela… viously held var… Supercomputing C… institute, he is also the Head of a specific s… on high productivity data processing as a p… Data Division. The given lectures in unive… Iceland, University of Applied Sciences of Co… University of Technology Aachen (RWTH A… "High Performance Computing and Big Dat… "Handling of large Datasets and Scientific ar… interests include "high productivity process… scientific computing applications.

Matthias Richerz… in scientific progr… Applied Sciences A… He is a Develop… tistical data minin… Group of High Pro… Supercomputing C… research interests … computing and mac… to enable scalable… quantities of scientific and engineering data.

Jón Atli Benediktsson (S'84–M'90–SM'99–F'04) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He is currently Rector and Professor of Electrical and Computer Engineering at the University of Iceland. His research interests include remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has authored extensively in those fields.

Dr. Benediktsson was the 2011–2012 President of the IEEE Geoscience and and Remote Sensing Society (GRSS) and has been on the GRSS AdCom since 2000. He was an Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) from 2003 to 2008 and has served as an Associate Editor of TGRS since 1999, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and the IEEE ACCESS since 2013. He is on the International Editorial Board of the *International Journal of Image and Data Fusion* and was the Chairman of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) from 2007 to 2010. He is a Cofounder of the biomedical start up company Oxymap (www.oxymap.com). He is a Fellow of SPIE. He is a member of the Association of Chartered Engineers in Iceland (VFI), Societas Scientiarum Islandica, and Tau Beta Pi. He was the recipient of the Stevan J. Krisiof Award from Purdue University in 1991 as Outstanding Graduate Student in remote sensing. In 1997, he was the recipient of the Icelandic Research Council's Outstanding Young Researcher Award; in 2000, he was granted the IEEE Third Millennium Medal; in 2004, he was a co-recipient of the University of Iceland's Technology Innovation Award; in 2006, he was the recipient of the yearly research award from the Engineering Research Institute, University of Iceland; and in 2007, he was the recipient of the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society. He is a co-recipient of the 2012 IEEE Transactions on Geoscience and Remote Sensing Paper Award. He was the recipient of the 2013 IEEE/VFI Electrical Engineer of the Year Award; and in 2013, he was a co-recipient of the IEEE GRSS Highest Impact Paper Award.

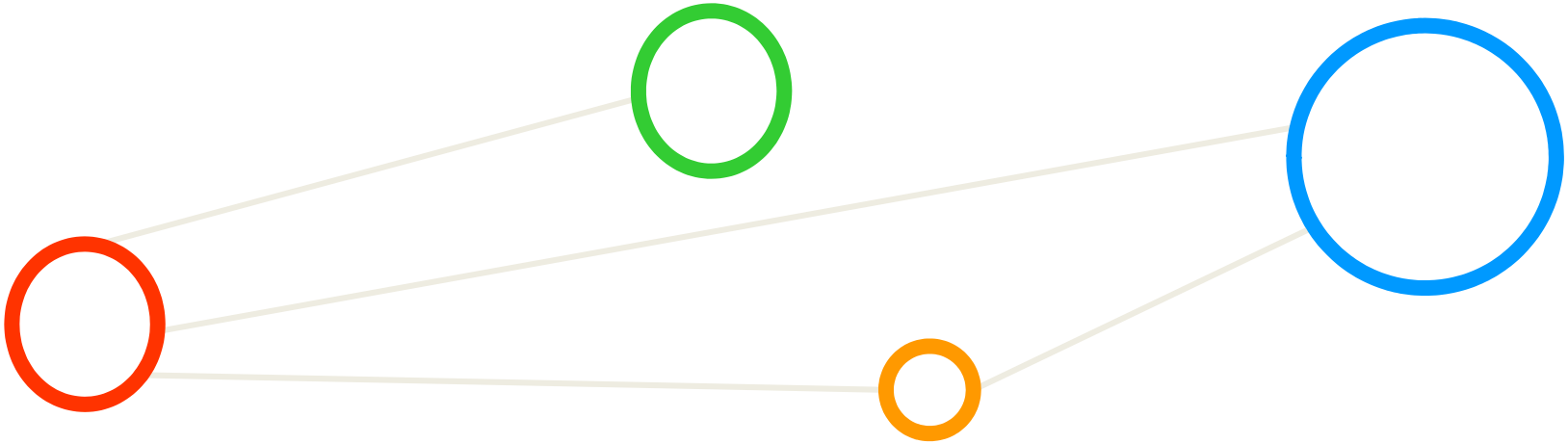
Antonio Plaza (M'05–SM'07–F'15) received the Computer Engineering degree and the M.Sc. and Ph.D. degrees in computer engineering from the University of Extremadura, Badajoz, Spain, in 1998, 2000, and 2002, respectively.

served as Associate Editor from 2007 to 2012.

Nächste „RDA" Schritte:

Datentypen per Type Registry definieren: Remote Sensing Daten, Resultate HPC Computer, SVM Datenformate, Feature Engineering Formate, etc.

Einladung zur Mitarbeit



Einladung zur Mitarbeit – Medizinbereich

- Reproducability
- Active Data Mar
- Big Data IG
- Biodiversity Dat
- Chemistry Resea
- Data Fabric IG
- Data for Develop
- Data Rescue IG
- …
- Domain Reposit
- Ethics and Socia
- Federated Ident
- Global Water Int

s WG
se WG
gistry WG
mework WG
vernance WG
WG
ion Registry
ty WG
ion and Terminology WG
gistries WG
and Trust WG
nteroperability WG

Wenige Gruppen im Bereich Medizin:

Bspw. Health Data Interest Group existiert

Einladung zur Mitarbeit:

Erzeugung neuer Gruppen nicht schwer

Bspw. TMF Dokumenten – Serie globaler?

Achtung: RDA gibt kein direktes Geld (Ausnahme 'Early Career' Unterstützung)

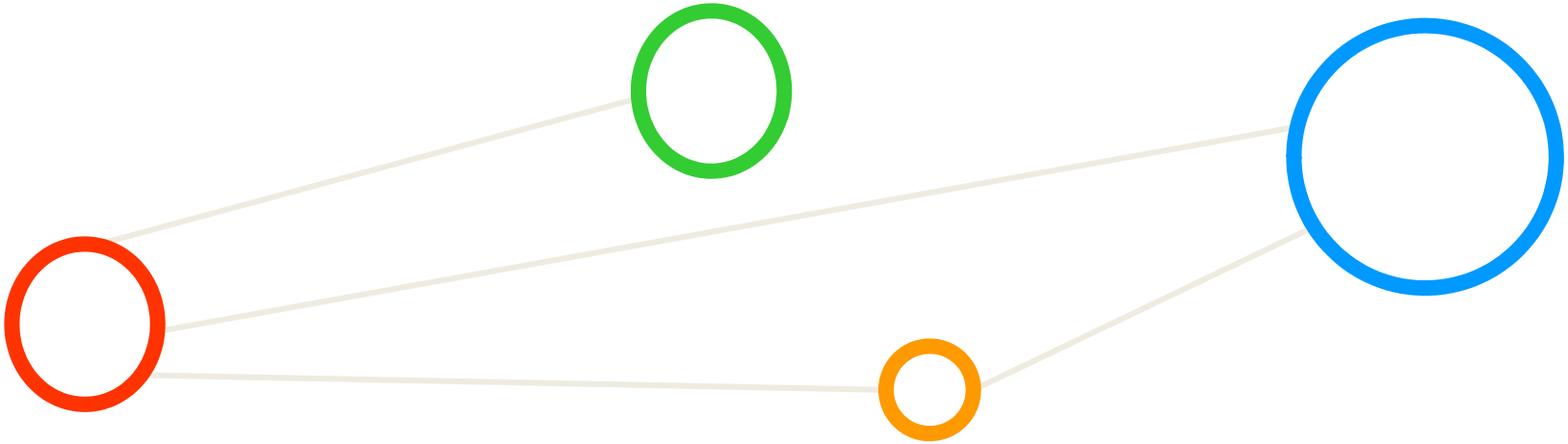
Konkrete erste Schritte:

Registrierung auf der RDA Webseite

Bei existierenden Gruppen anmelden

Neue Gruppen bei Tagungen vorschlagen

Zusammenfassung



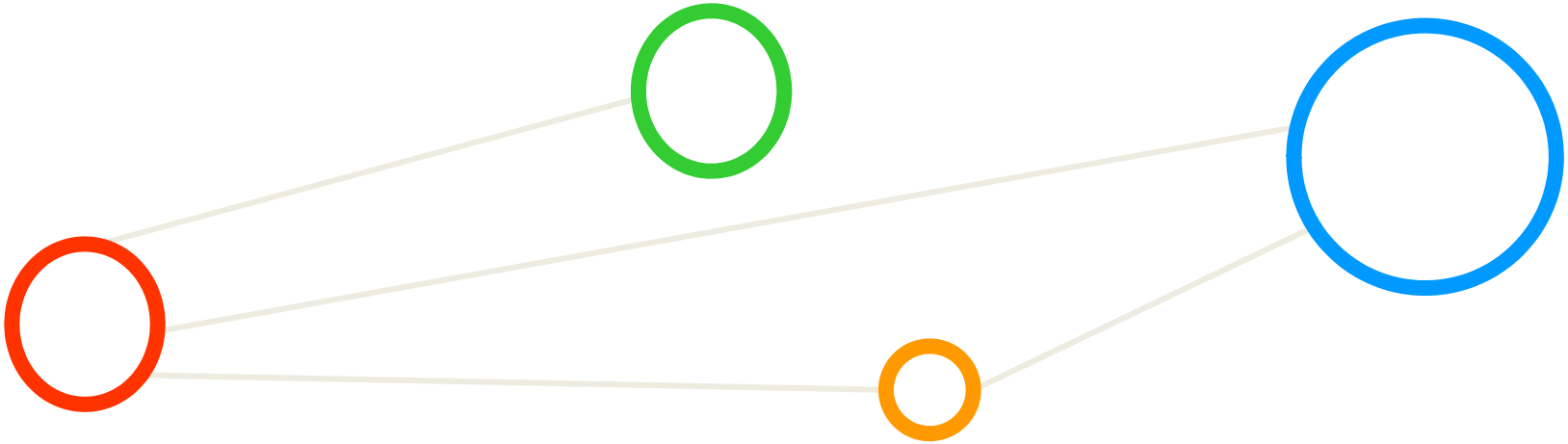
Zusammenfassung

- Teilnahme an RDA für jeden offen
  - Arbeit wird vorangebracht durch aktive(!) Teilnehmer
  - Individuelle Teilnehmer sowie Organisationen können Mitglied werden
  - Erster Schritt : Registrierung auf der RDA Webseite & Gruppen suchen
  - Möglichkeit durchaus (relativ) einfach neue Gruppen zu erstellen

- Wichtig: Resultate & Umsetzung
  - Resultate werden von Arbeits-Gruppen erzeugt und sind offen verfügbar
  - Initiale Ergebnisse von Arbeitsgruppen sind nun verfügbar

RDA funktioniert langfristig nur durch Umsetzung der Empfehlungen und „Best Practices“

Referenzen



Kurzer Hinweis: Smart Data Innovation Lab

ABB AG Forschungszentrum
BASF SE
Bayer Technology Services GmbH
Carl von Ossietzky Universität Oldenburg, Abt. Wirtschaftsinformatik I / VLBA
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Deutsches Krebsforschungszentrum
Blue Yonder GmbH & Co. KG
BOGEN Electronic GmbH
Robert Bosch GmbH
Fraunhofer Institut Intelligente Analyse- und Informationssysteme IAIS
Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO
Forschungszentrum Jülich GmbH
CAS Software AG
EBID Service AG
echobot Media Technologies GmbH
econda GmbH
EnBW Energie Baden-Württemberg AG
Hitachi Data Systems GmbH
FZI Forschungszentrum Informatik
Karlsruher Institut für Technologie (KIT)
RWTH Aachen, Joint Research Center for Computational Biomedicine
IBM Deutschland GmbH
Infineon Technologies AG
LeserAuskunft GmbH
TU Berlin, Fachgebiet Datenbanksysteme und Informationsmanagement
Microsoft Deutschland GmbH
PTV GROUP - PTV Planung Transport Verkehr AG
SAP AG
Siemens Aktiengesellschaft
SOFTWARE AG
Stratosphere
TRUMPF Werkzeugmaschinen GmbH + Co. KG
Volkswagen AG

Smart Data Innovation Lab

[7] SDIL Website (Online)

Kern: Datenanalyse Wirtschaft&Wissenschaft

SDIL Medizin Bereich: ~20 Mitglieder, 3 Probleme

Referenzen

[1] RDA Website (Online): https://rd-alliance.org/

[2] RDA DFT WG (Online): https://rd-alliance.org/group/data-foundation-and-terminology-wg.html

[3] RDA DTR WG (Online): https://www.rd-alliance.org/group/data-type-registries-wg.html

[4] G. Cavallaro, M. Riedel, M. Richerzhagen, J.A. Benediktsson, A. Plaza ***On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods,*** IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Issue 99, pp. 1-13, 2015

[5] EUDAT B2Share Service zur Archivierung (Online): https://b2share.eudat.eu/

[6] piSVM – Parallel Support Vector Machine (Online): http://pisvm.sourceforge.net/index.html

[7] SDIL Website (Online): http://www.sdil.de/

RESEARCH DATA ALLIANCE

DANKE

Folien sind in Kürze erhältlich auf:

http://www.morrisriedel.de

research data sharing without barriers

rd-alliance.org