RESEARCH DATA ALLIANCE

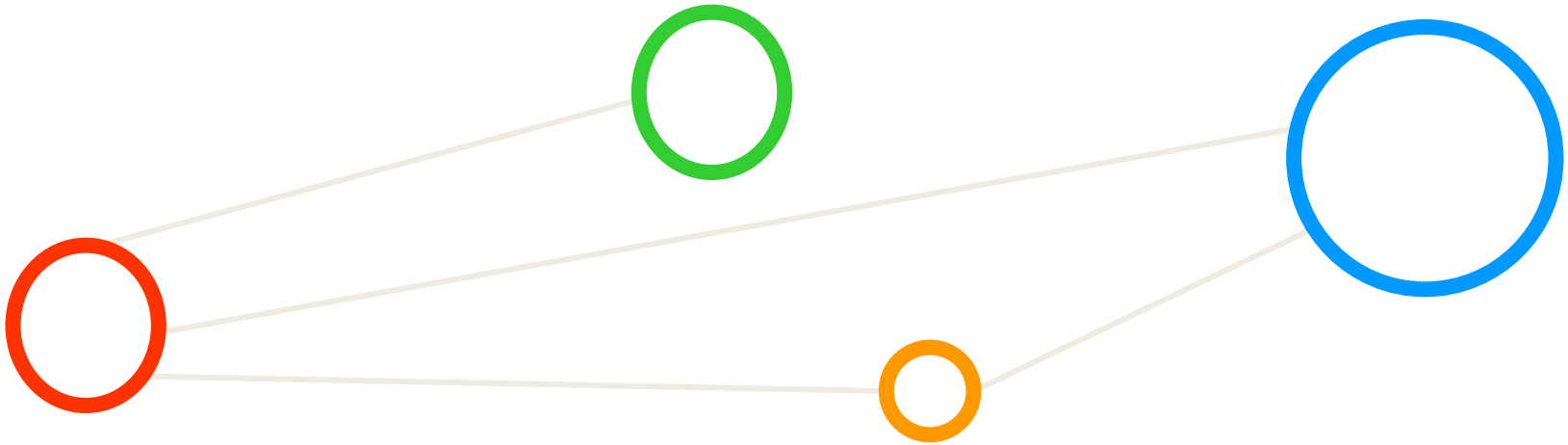
Eine Weltweite Initiative

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

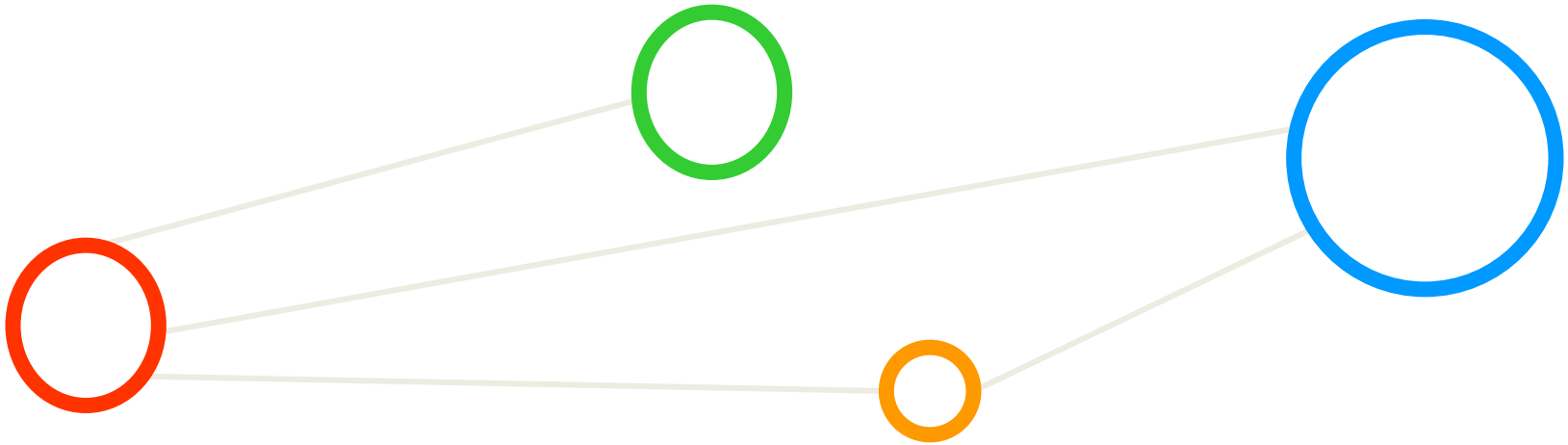
Gliederung



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

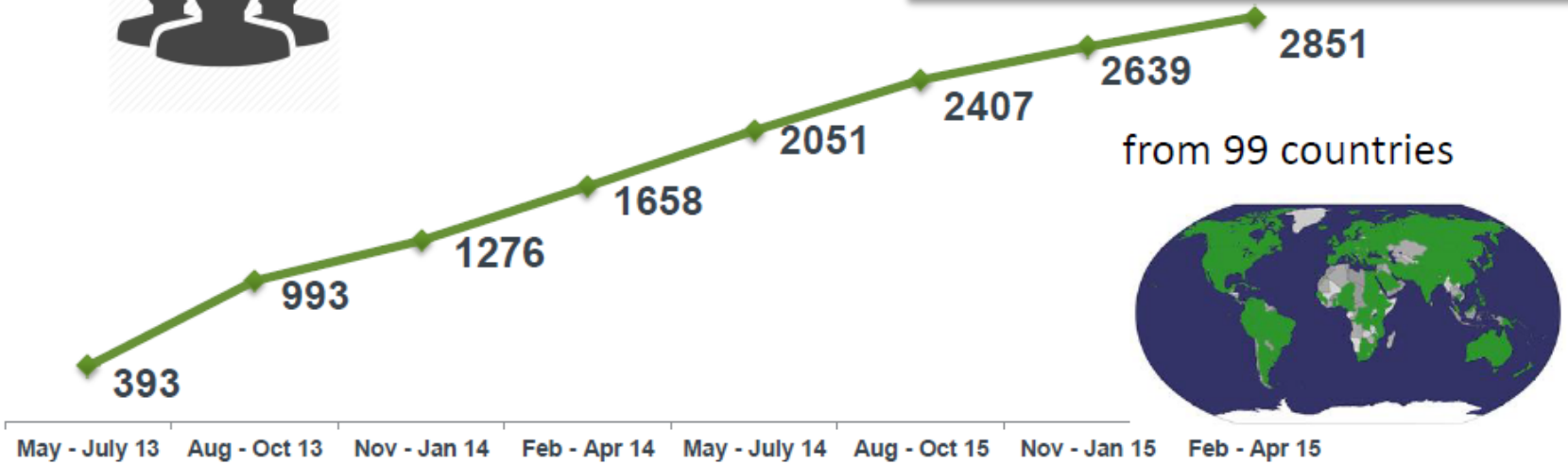
- „RDA Vision“ verwirklichen
  - Weltweit: Überwindung von Ländergrenzen
  - Forscher können frei verfügbare Daten austauschen
  - Austausch über Technologiegrenzen und Disziplinen
- „RDA Mission“
  - Aufbau von „sozialen und technischen Brücken“ für den Datenaustausch
- Angestrebte Resultate
  - Erstellen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

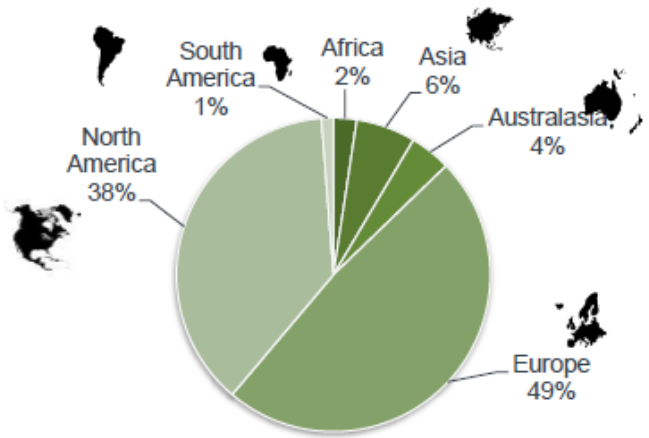
[1] RDA Website (Online)

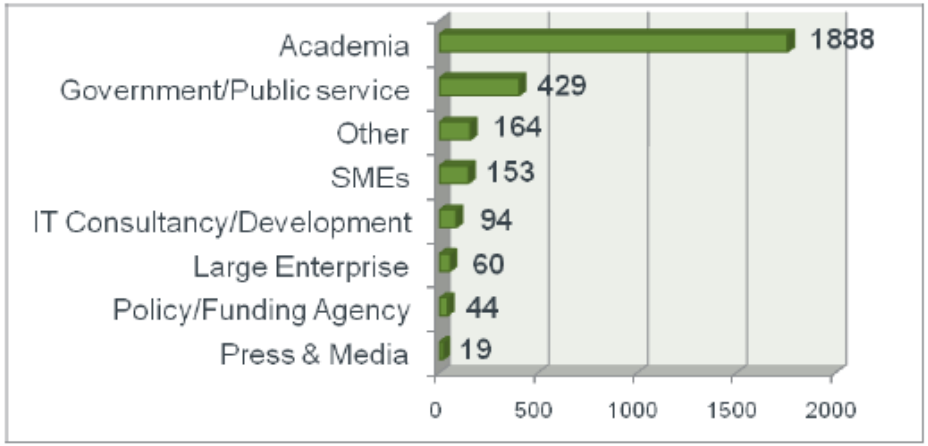
Wichtig: RDA fördert zwar den Datenaustausch – sammelt aber selbst keine Daten ein!

Internationale Mitglieder

Totale Anzahl Mitglieder heute: ~3000







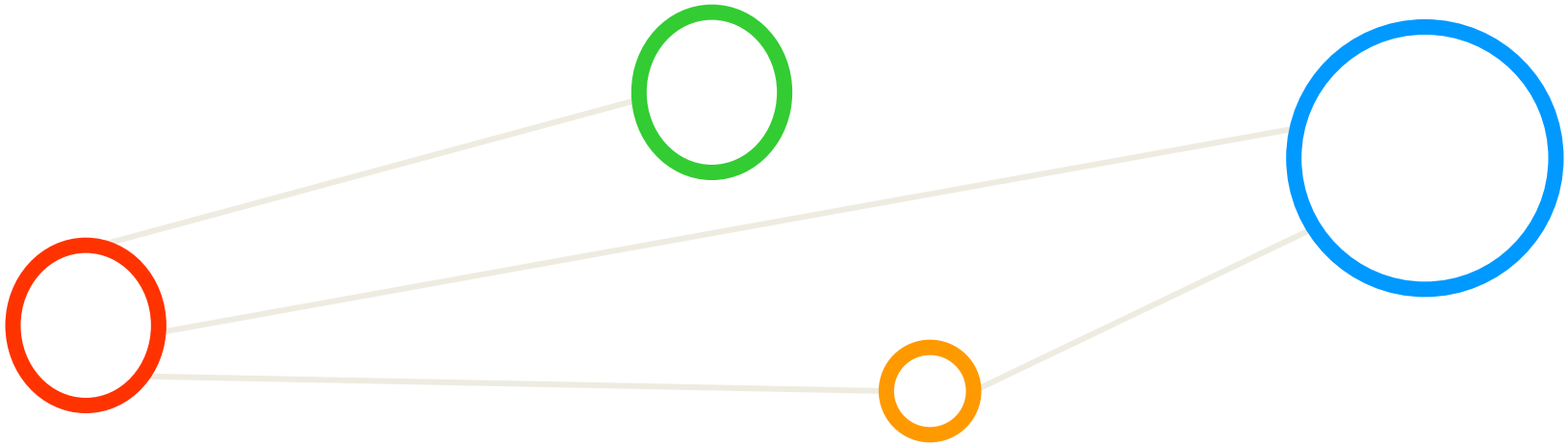
Mitglieder Organisationen

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

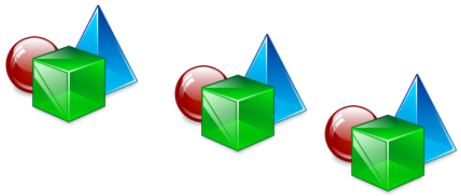
- Interessen – Gruppen (engl. ‘interest groups‘)
  - Ziel: Identifikation gemeinsamer Schwerpunktthemen
  - Zeitlich unbefristet, oft aus ‘Communities‘ heraus
  - Keine direkten Resultate angestrebt

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

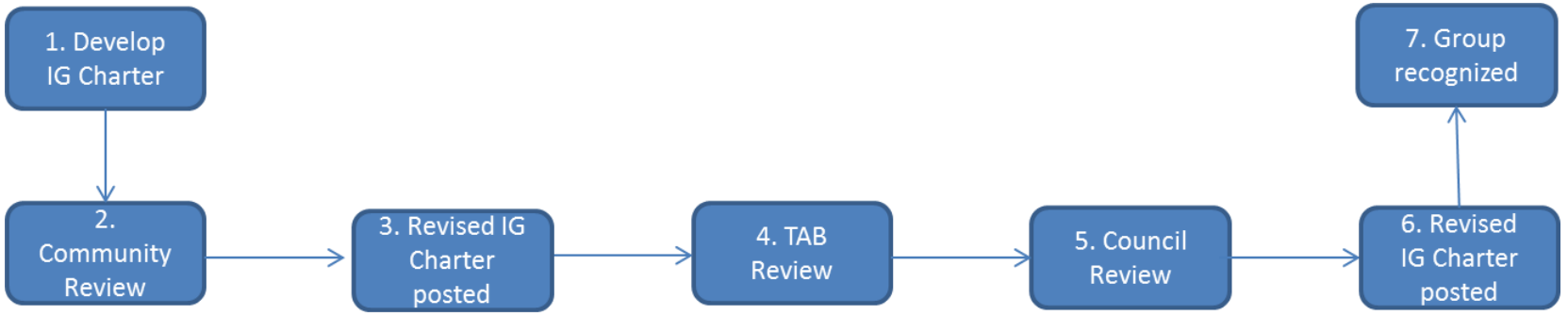
- Reproducability IG
- Active Data Management Plans IG
- Big Data IG
- Biodiversity Data Integration IG
- Chemistry Research Data IG
- Data Fabric IG
- Data for Development IG
- Data Rescue IG
- …
- Domain Repositories IG
- Ethics and Social Aspects of Data IG
- Federated Identity Management IG
- Global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

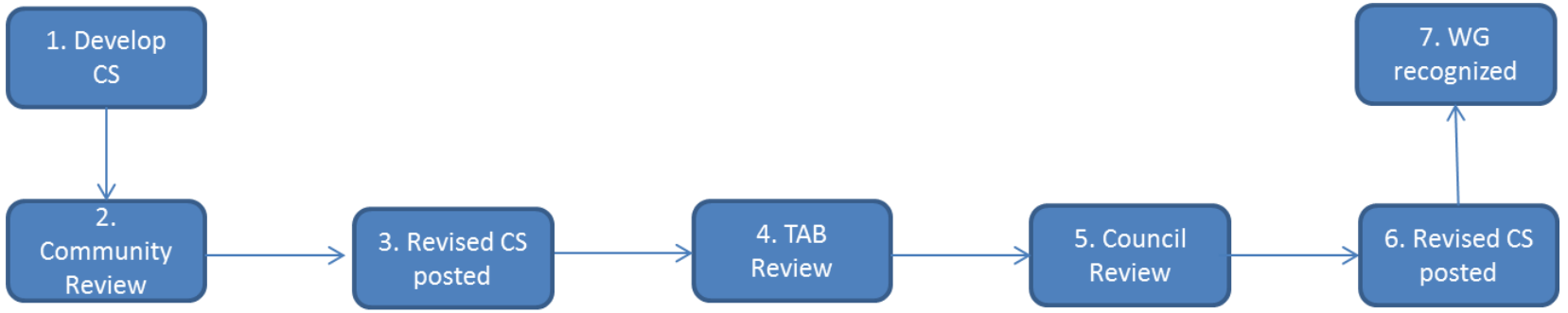
Erstellung neuer Gruppen

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

Prozess für Interessen – Gruppen



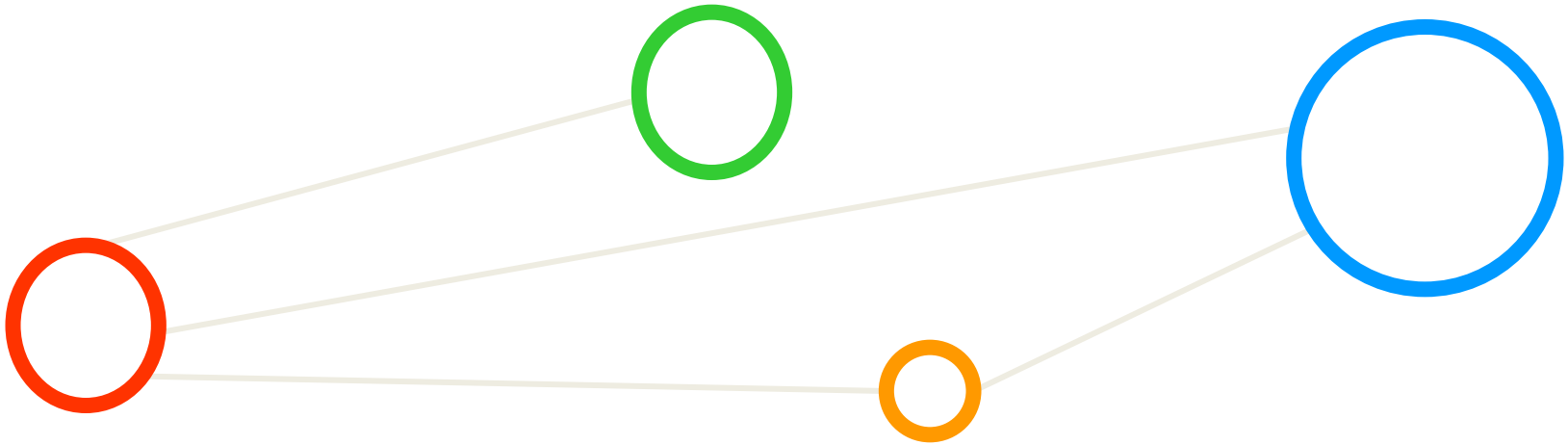
Prozess für Arbeits – Gruppen



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.

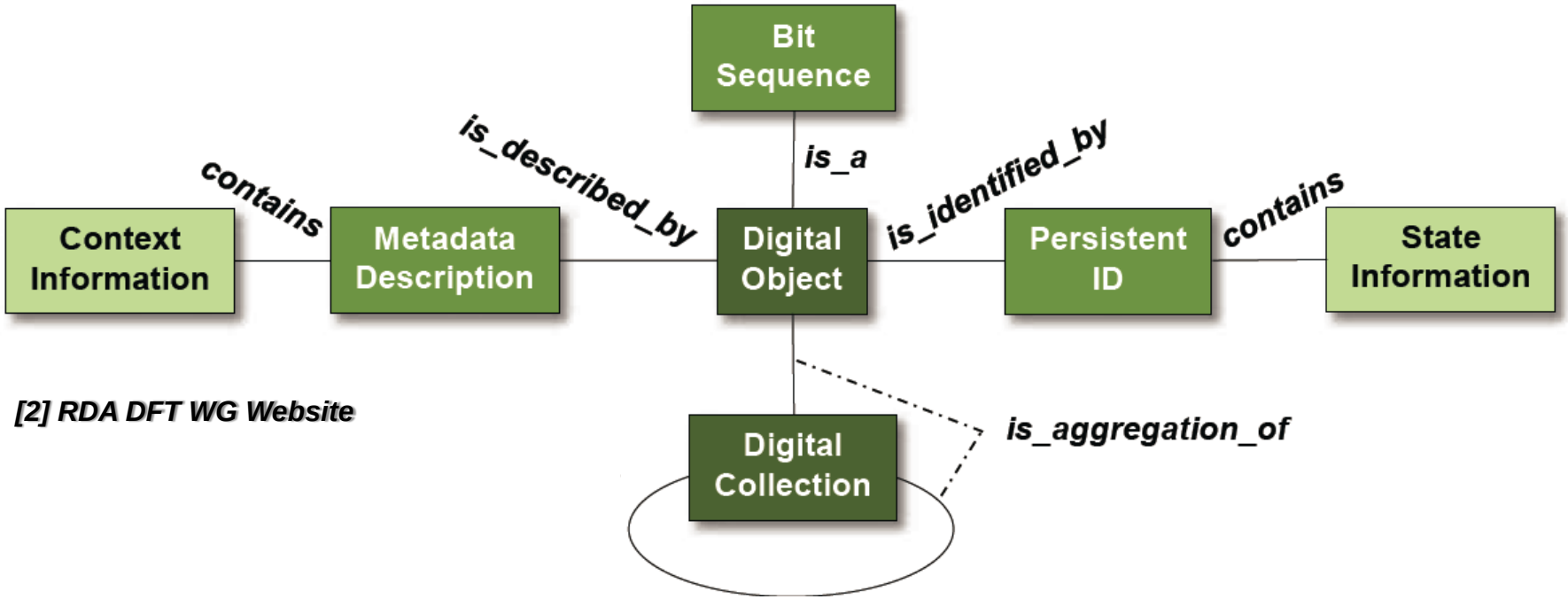
Hadoop is very powerful but it is important to understand that it requires an underlying computing infrastructure that is different from a usual desktop computer or simple laptop. It is thus not a solution like R or Matlab that you simply run on your own computer and therefore users of this tool usually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

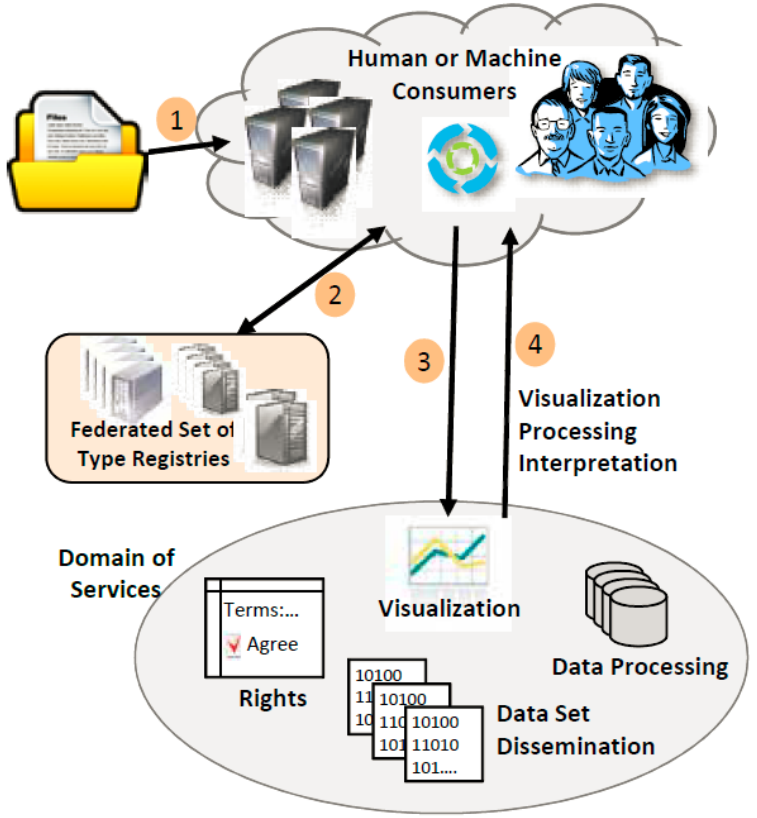
- Data Foundation and Terminology WG Resultat
  - Ziel: Einigung auf ein sehr einfaches Datenmodell wäre prima in der Praxis
  - Diskussion der Frage: „Was ist ein digitales Objekt nun genau?“
  - Enthält Definitionen: Persistent Identifier (PID), Digital Object, Metadata, …



[2] RDA DFT WG Website

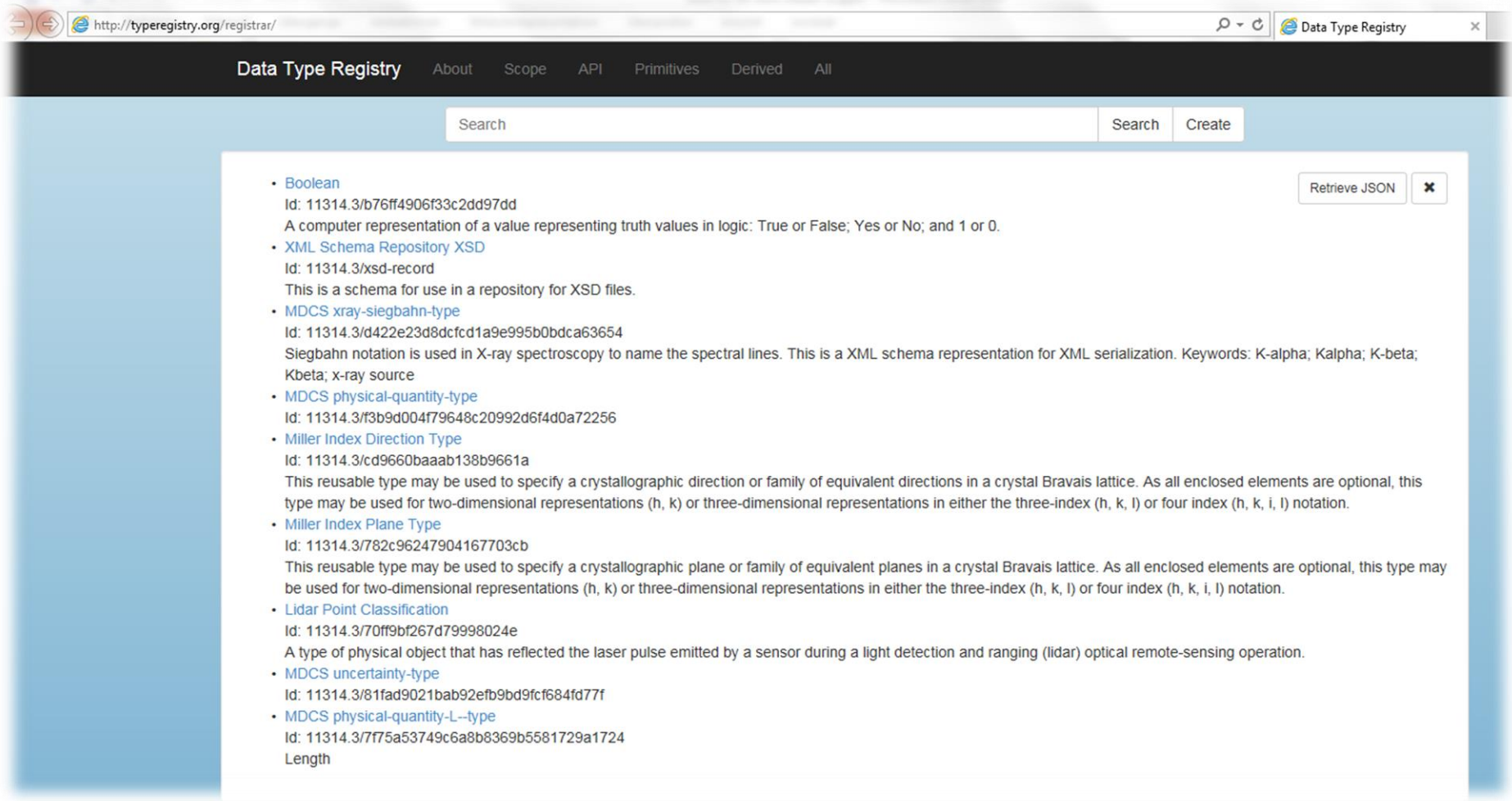
Erste Resultate – Wissenschaftliche Datentypen

- 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/



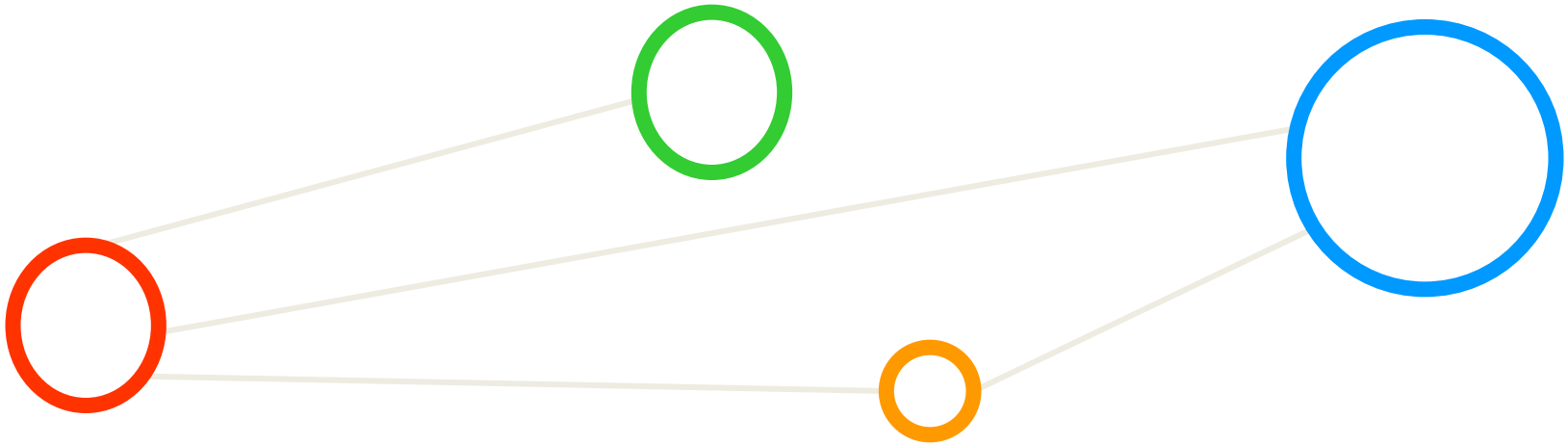
[3] RDA DTR WG Website

Erste Resultate – Wissenschaftliche Datentypen



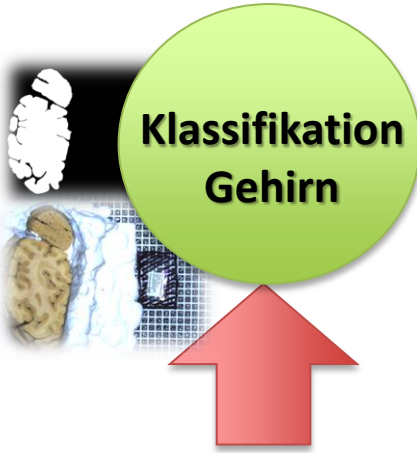
[3] RDA DTR WG Website

Umsetzungsbeispiele



Umsetzung PID: Archivierung von Forschungsdaten

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



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

„Peer-Review“

„Best Practices“

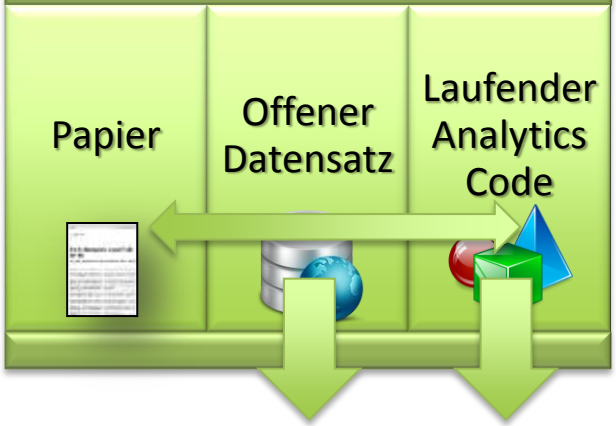
Parallel Support Vector Machines (SVM)

Klassifikation Erdboden-Bedeckung

HPC/MPI, Map-Reduce & GPGPUs

Community-basierte Empfehlungen

„Reference Data Analytics“ zur Wiederverwendung



B2SHARE Store and Share Research Data

[5] EUDAT B2SHARE

πSVM

[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

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

The record appears in these collections:
Generic

Metadata

Daten und Skripte archiviert und vom Papier per PID referenziert ☐ entspricht Definition RDA „Digital Object“ für die Papierdaten

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

Metadata

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

PID handle

Umsetzung PID: Archivierung von Forschungsdaten

[62] F. Garcia-Vilchez *et al.*, "On the impact of lossy compression on hyperspectral 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)

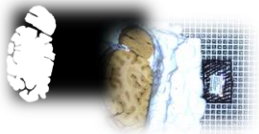
B2SHARE
Store and Share Research Data

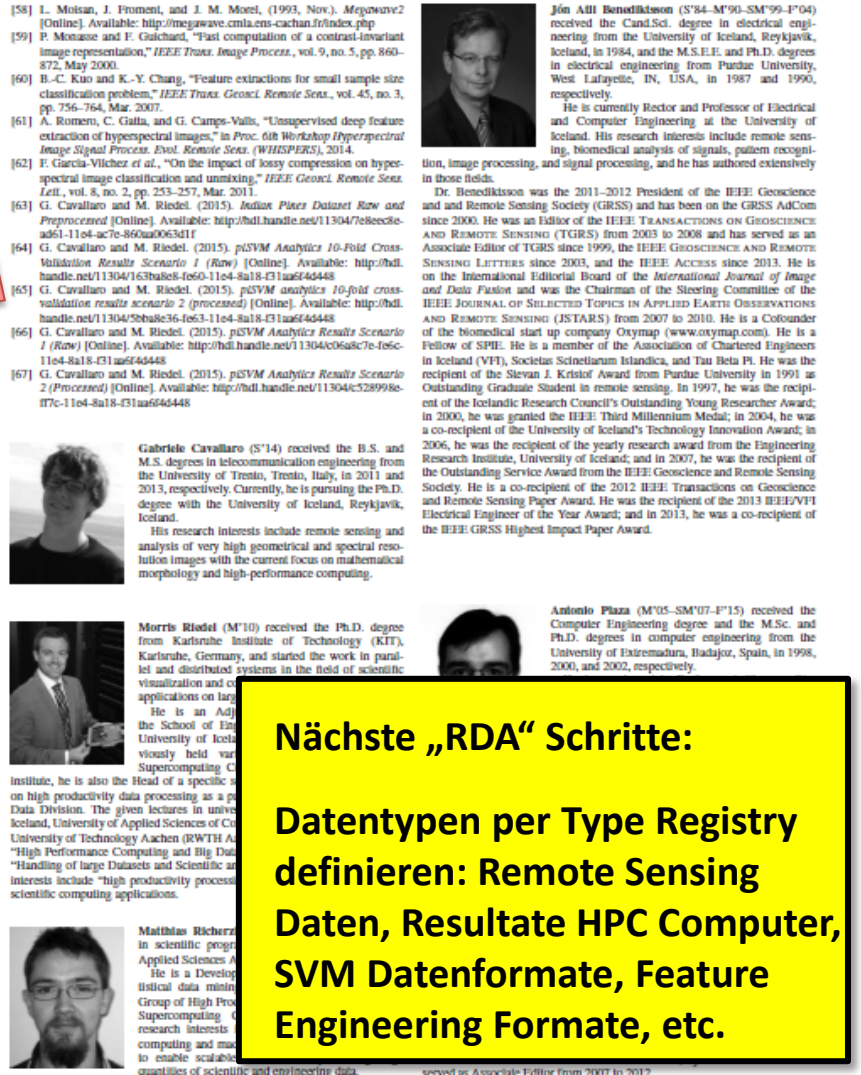
[5] EUDAT B2SHARE

more reputation for authors

enables data-sharing

enables reproducability





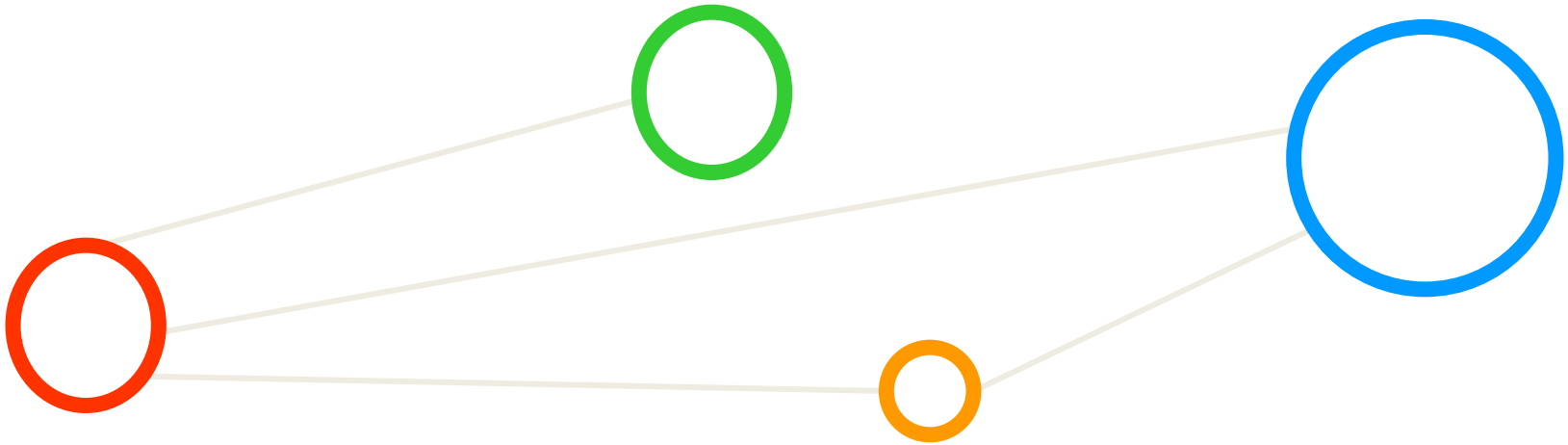
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

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

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 Data
- Chemistry Resea
- Data Fabric IG
- Data for Develop
- Data Rescue IG
- …
- Domain Reposit
- Ethics and Socia
- Federated Ident
- Global Water Int

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)

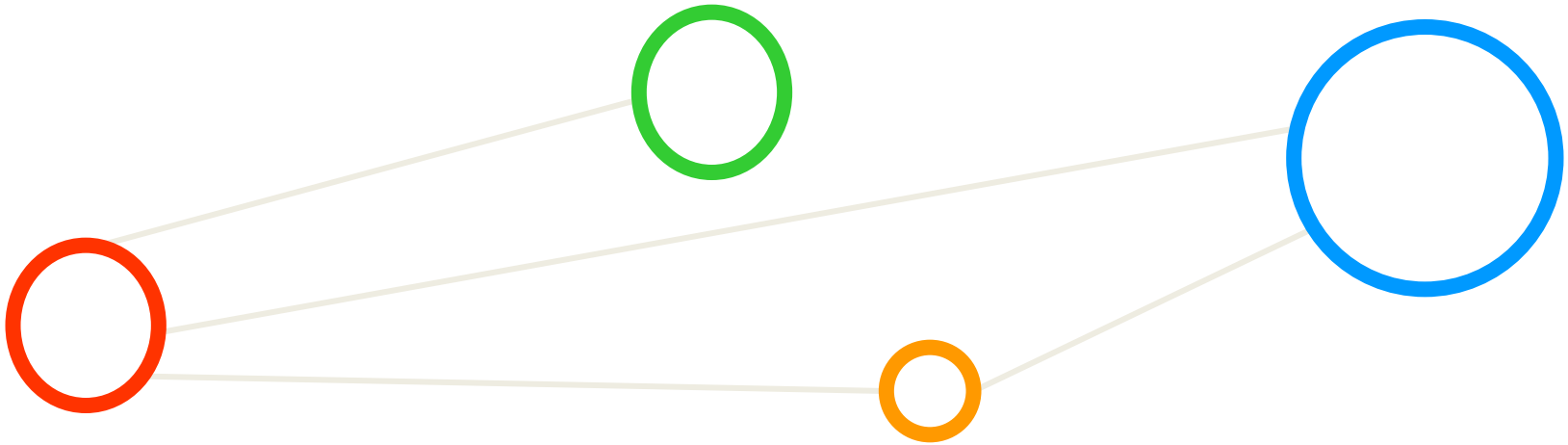
Konkrete erste Schritte:

Registrierung auf der RDA Webseite

Bei existierenden Gruppen anmelden

Neue Gruppen bei Tagungen vorschlagen

Zusammenfassung



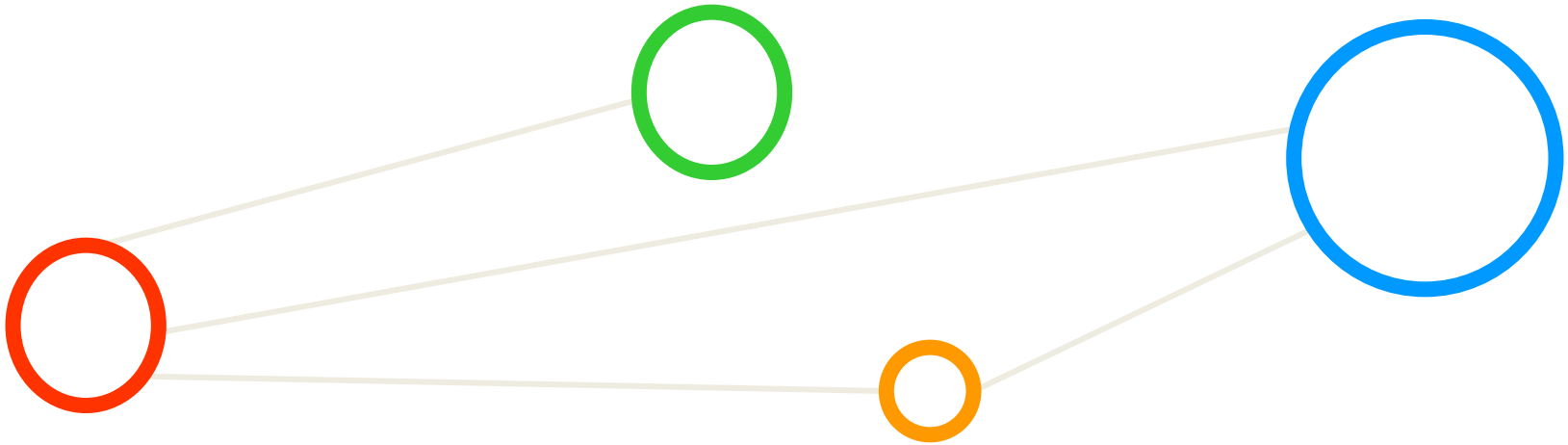
Zusammenfassung

- 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

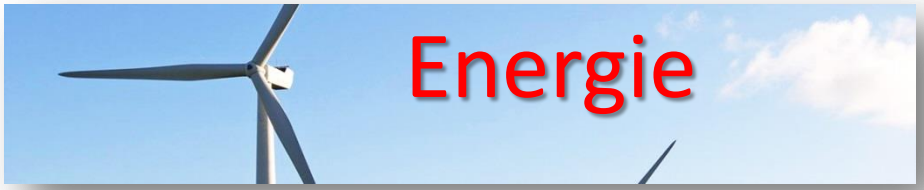


ABB AG Forschungszentrum
BASF SE
Bayer Technology Services GmbH
Carl von Ossietzky Universität Oldenburg, Abt. Wirtschaftsinformatik I / VLBA
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Deutsches Krebsforschungszentrum
Blue Yonder GmbH & Co. KG
BOGEN Electronic GmbH
Robert Bosch GmbH
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/

DANKE

Folien sind in Kürze erhältlich auf:

http://www.morrisriedel.de

research data sharing without barriers

rd-alliance.org