



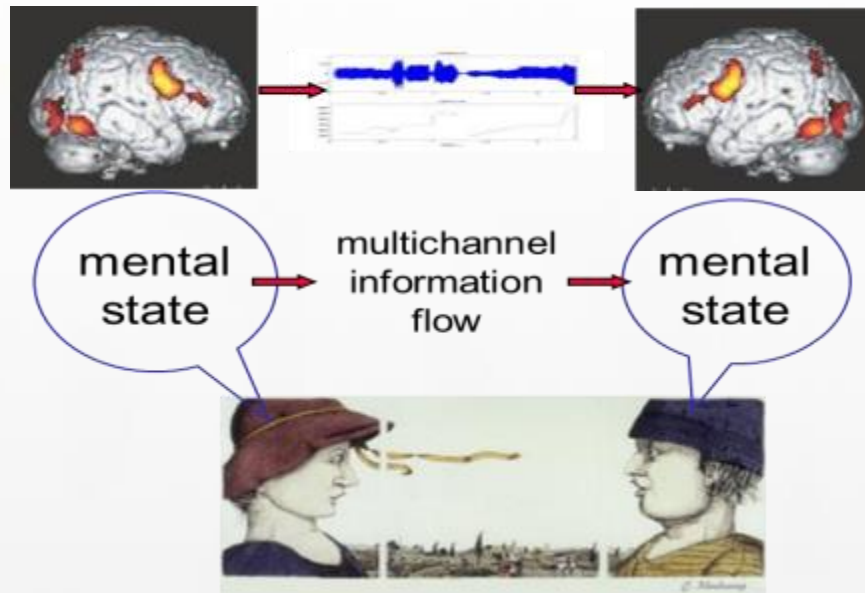
RESEARCH DATA ALLIANCE
EUROPE

RDA – Daten als Teil der Wissenschaftskultur

Peter Wittenburg
Max Planck Institut für Psycholinguistik
RDA E Scientific Coordinator
RDA TAB Member

Psycholinguistik

Stimuli
↓
einfache
Reaktionen



H. Billing
MPG IT Rat
1976

Billing: So lieber Herr Levelt - jetzt erzählen Sie uns mal, wozu denn die Psycholinguisten Computer haben wollen. Ich dachte ...

Levelt: Lieber Herr Billing - wir wollen das Sprachverhalten des Gehirns ausmessen.



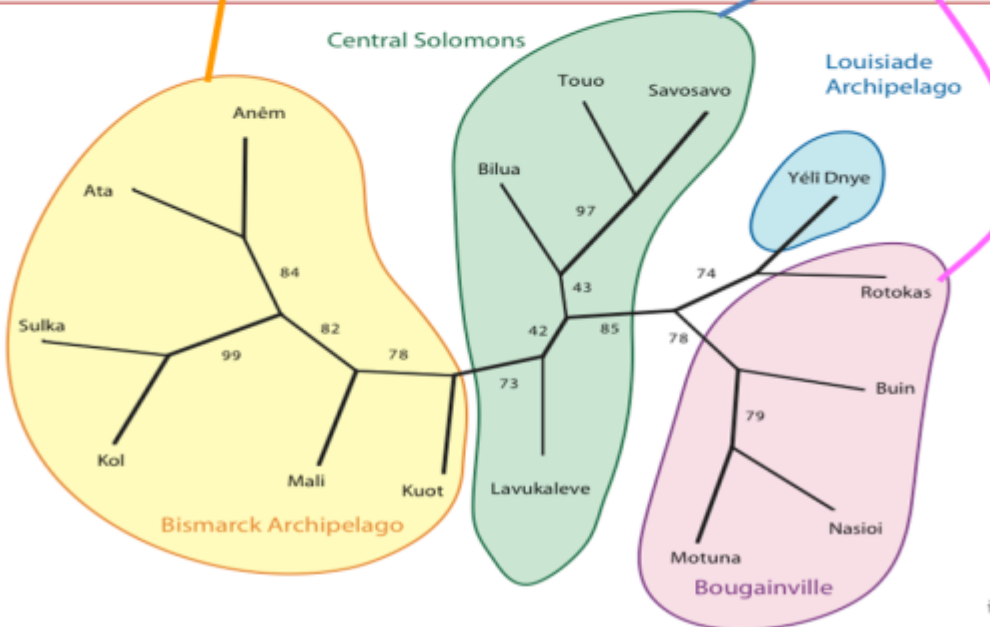
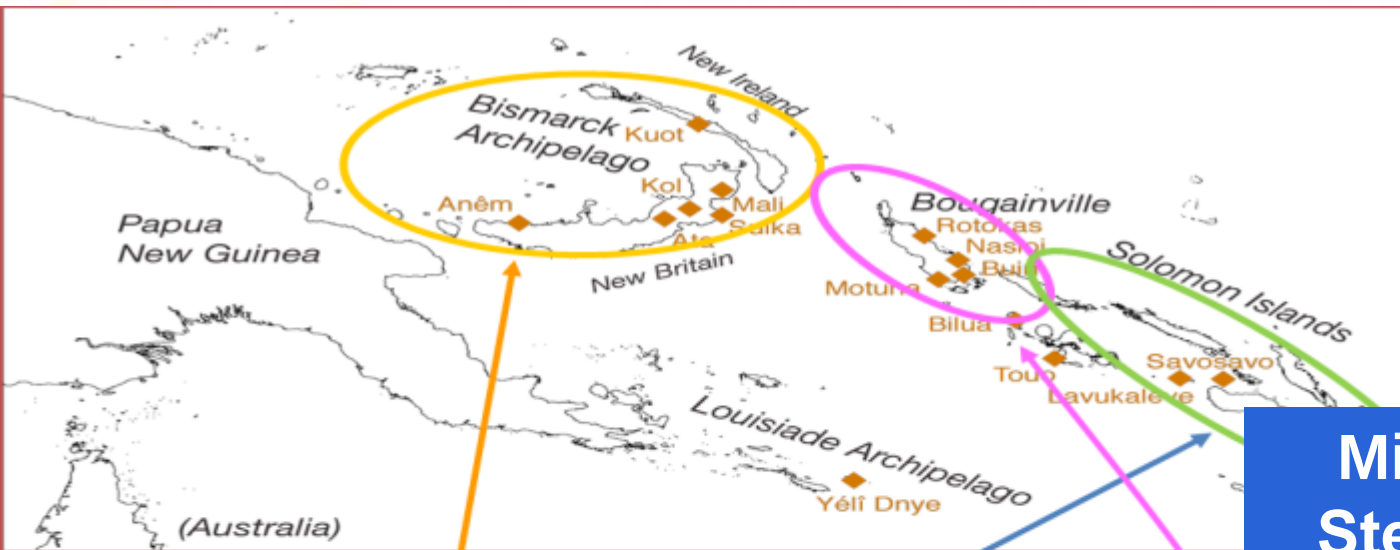
P. Levelt
Nijmegen
1976

Daten-Beispiel I



**Hans Rosling's
Gapminder:
von fragmentierten
Daten zu
Einsichten**

Daten-Beispiel II

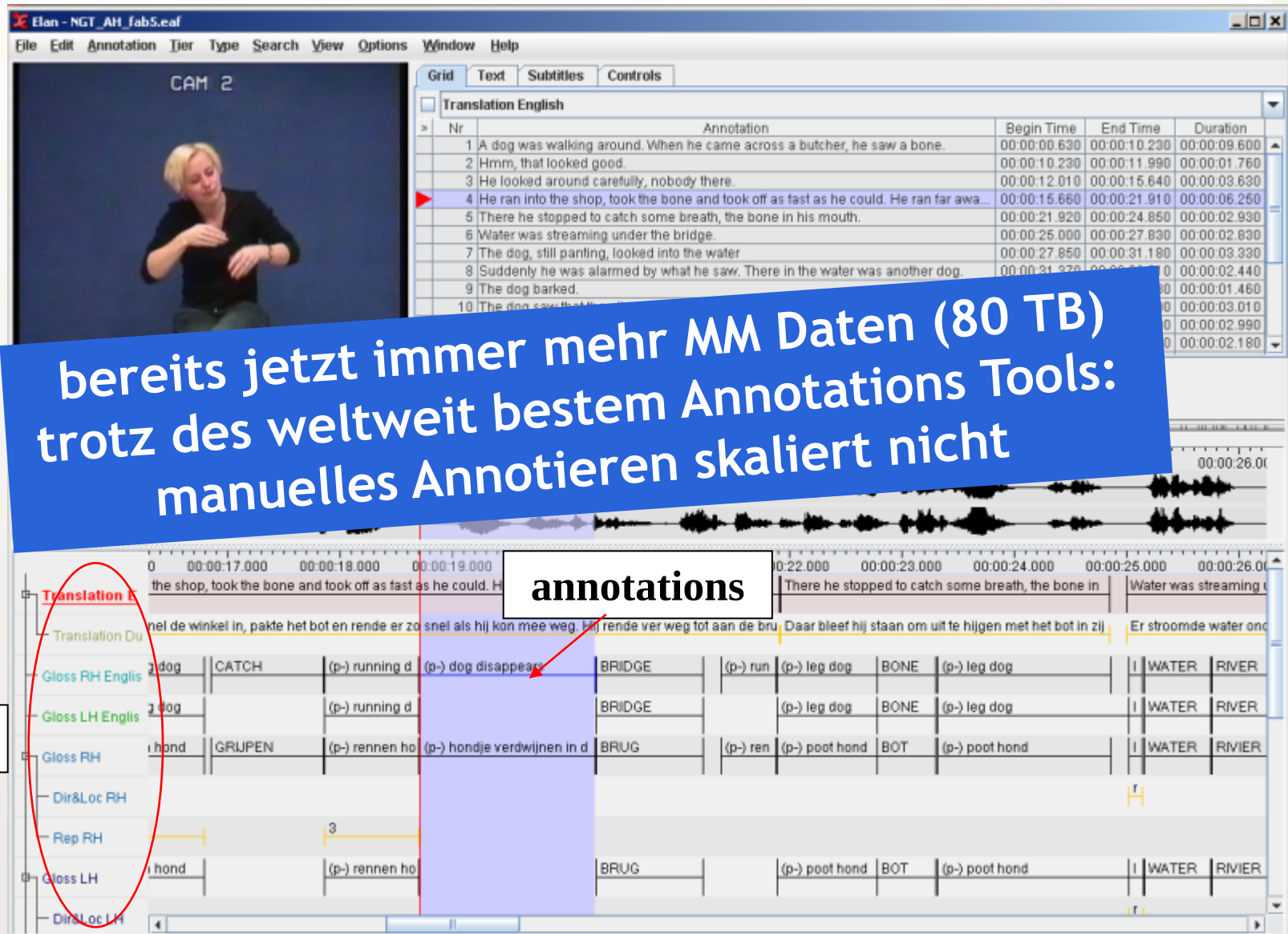


**Michael Dunn &
Steve Levinson's
Wurzeln der
Sprachen:
von fragmentierten
Daten zu
Einsichten**

viele andere Beispiele aber ...

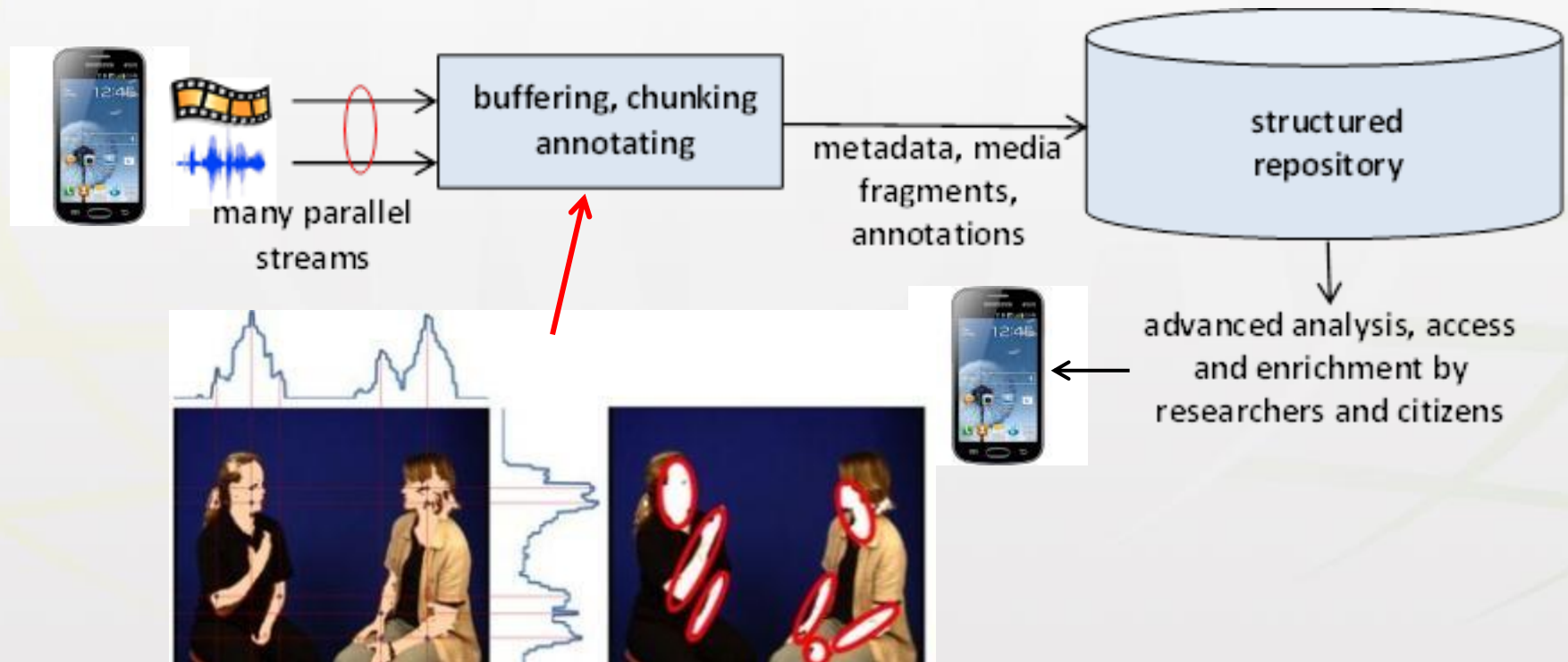
- Daten werden oftmals mit einem erheblichen Aufwand manuell integriert
- wird dies skalieren in einer Zeit, in der alle Wissenschaften immer mehr und komplexere Daten erzeugen?
- UND – wer kann sich diese Art der Wissenschaft bei einem derartigen Aufwand erlauben?
- noch ein weiteres Beispiel aus der (veränderten) Welt der Geisteswissenschaft

manuelles Annotieren



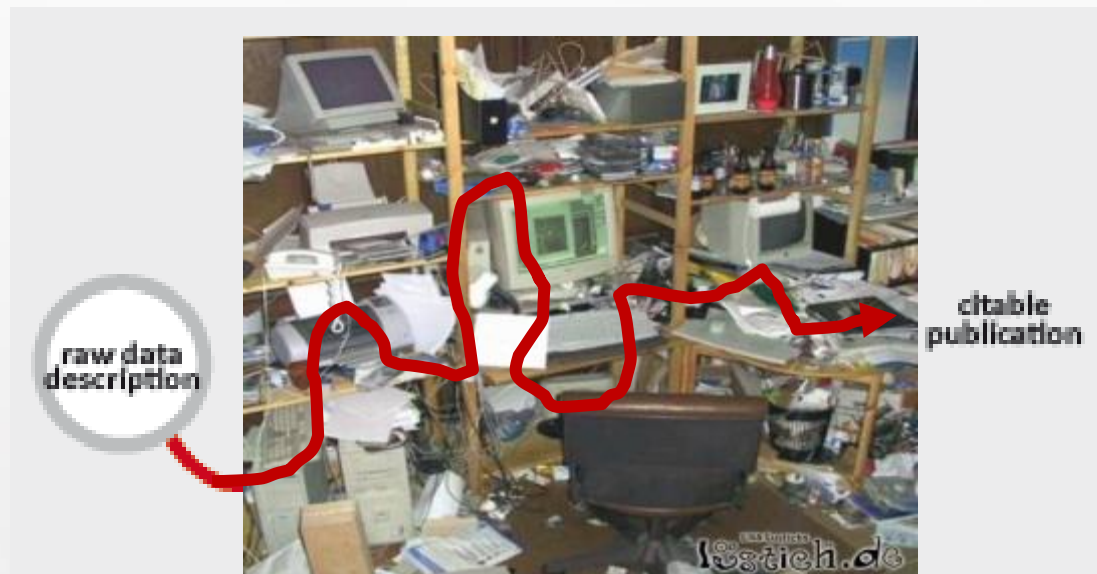
Crowd Sourcing: noch mehr Daten

- massives CS über mobile Geräte: viele Versuchspersonen und MD mit vielfältigen Sensor-Typen
- 10 min * 100 P/Tag multimedia Aufnahmen (H.264) = **100 GB/Tag**
- ohne eine Maschinerie zur Reduktion, Annotation und zum Daten-Management hoffnungslos

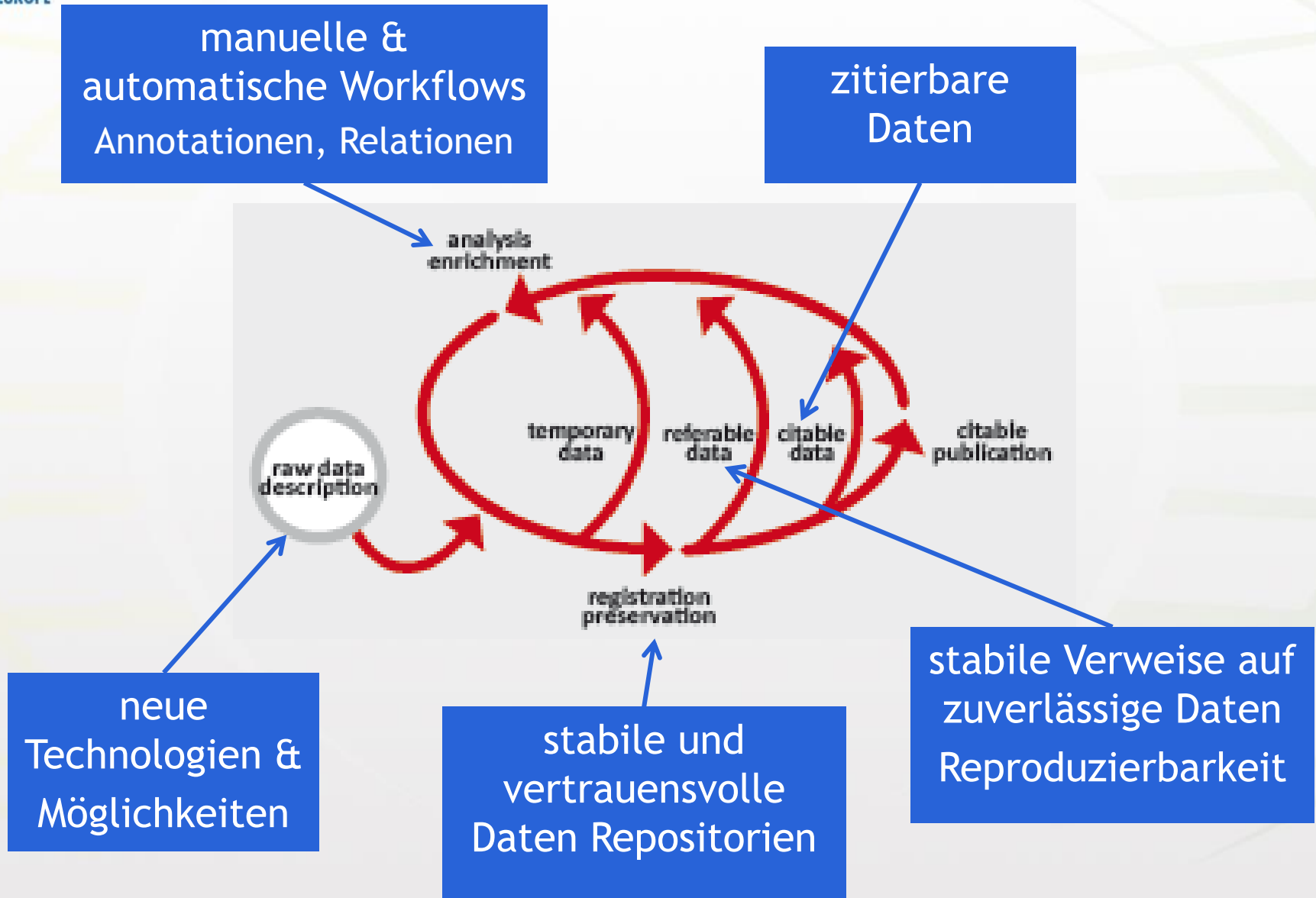


über Daten sprach man nicht viel

über verschlungene Pfade zur Publikation



Daten sind ein Thema



sind wir dem Ansturm gewachsen?

- Daten-Management, -Auffinden & -Zugang ist ineffizient und kostet zu viel
 - ausgezeichnete Wissenschaftler “vergäuden” 50+ % ihrer Zeit mit Aspekten des Daten Managements und des Daten-Zugangs
 - alte Strukturen wie File Systeme werden verwendet, um die Organisation der Daten vorzunehmen
 - Geldgeber sehen DM/DA Anträge von allen Disziplinen – viele Steuergelder werden immer wieder für die gleiche Komponenten ausgegeben
- Wir sind dem Ansturm der großen Mengen und der Komplexität nicht gewachsen!

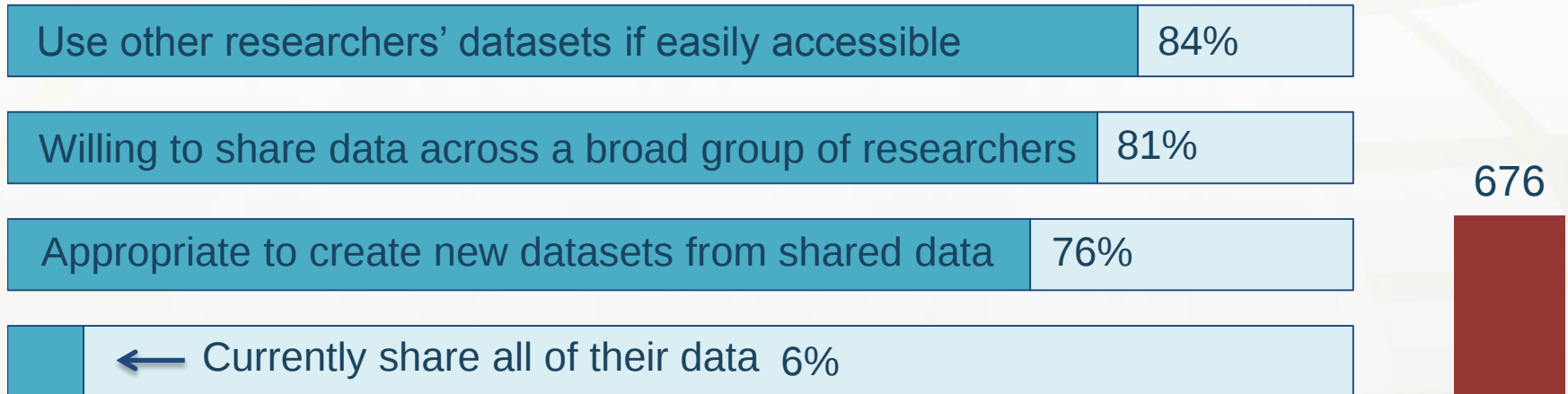
viele Interviews

- viele Interviews, um die Daten-Landschaft zu verstehen
 - RDA Europe Interviews (27)
 - Modell Studien in RDA DFT WG (22)
 - EUDAT & Radieschen Interviews (17)
- Haupt-Schlussfolgerungen
 - Daten-Architekturen und Organisationen sind alle unterschiedlich
 - große Heterogenität bezüglich der verwendeten Software-Pakete
 - natürlich: die Datenmengen steigen, die Komplexität durch vielfältige Formate und Referenzen nimmt zu, alte Lösungen gehen nicht mehr
 - zusätzlich:
 - keine Auffindbarkeit
 - Metadaten sind ein Albtraum, also ist Wiederverwendung kaum möglich
 - einfacher Zugang zu Daten ist kaum möglich
 - Reproduzierbarkeit der Resultate ist kaum gegeben

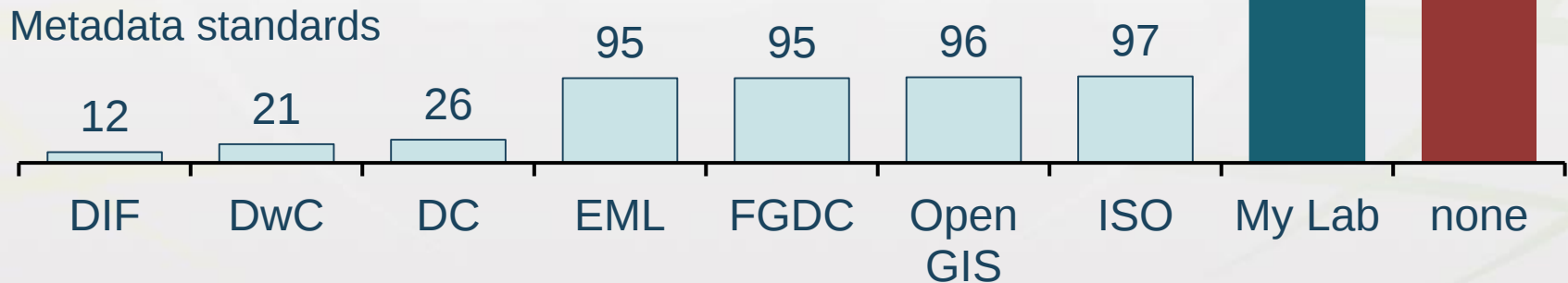
einige führende Wissenschaftler

- RDA – MPG Workshop zu wiss. Daten
- 17 führende Wissenschaftler (EU, Cross-Disz.)
- einige wichtige Punkte:
 - das Umgehen mit Menge und Komplexität wird aufwendiger
 - cross-disziplinärer Daten-Zugriff und –Kombination ist Realität
 - es fehlen cross-disziplinäre Absprachen (MD, Formate, etc.)
 - Vertrauen in und Zitierbarkeit von Daten ist wichtig
 - automatische Workflows werden immer bedeutender
 - wollen eine funktionierende/persistente Infrastruktur
 - anerkennen, dass Infrastrukturen Spezifikationen bedürfen
 - trotz begründeter Skepsis: RDA ist sinnvoll
 - Lösungsfindung muss bottom-up organisiert sein

Wissenschaftler wollen Daten teilen



... aber wissen oft nicht wie!



slide from Bill Michener, DataONE

Sorgen der EU Cluster Projekte

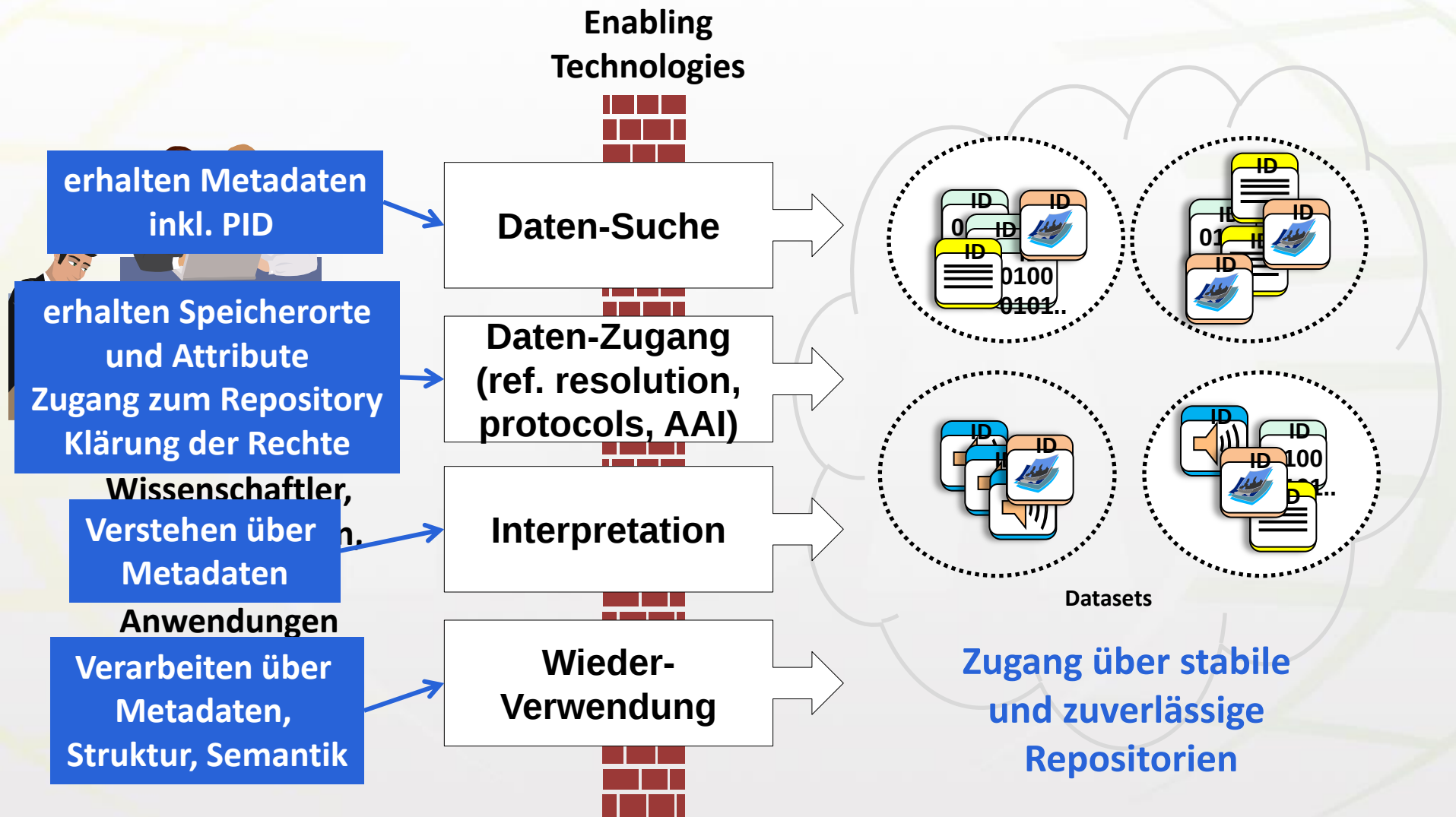
Identität

	CRISP	ENVRI	DASISH	BioMed
Data identity				
Data identity continuum				
Software identity				
Concept identity				
User identity management				
Common data standards and formats				
Service discovery				
Service market places				
Integrated data access and discovery				
Data storage facilities				
Data curation				
Privacy and security				
Dynamic data management				
User Community Body				
Semantic annotations and bridging				
Reference models				
Education & training				

Finden

Persistenz
Kuration

Identifizieren einiger Komponenten



graphic from Larry Lannom, CNRI

ein kurzer Blick zurück - 2011

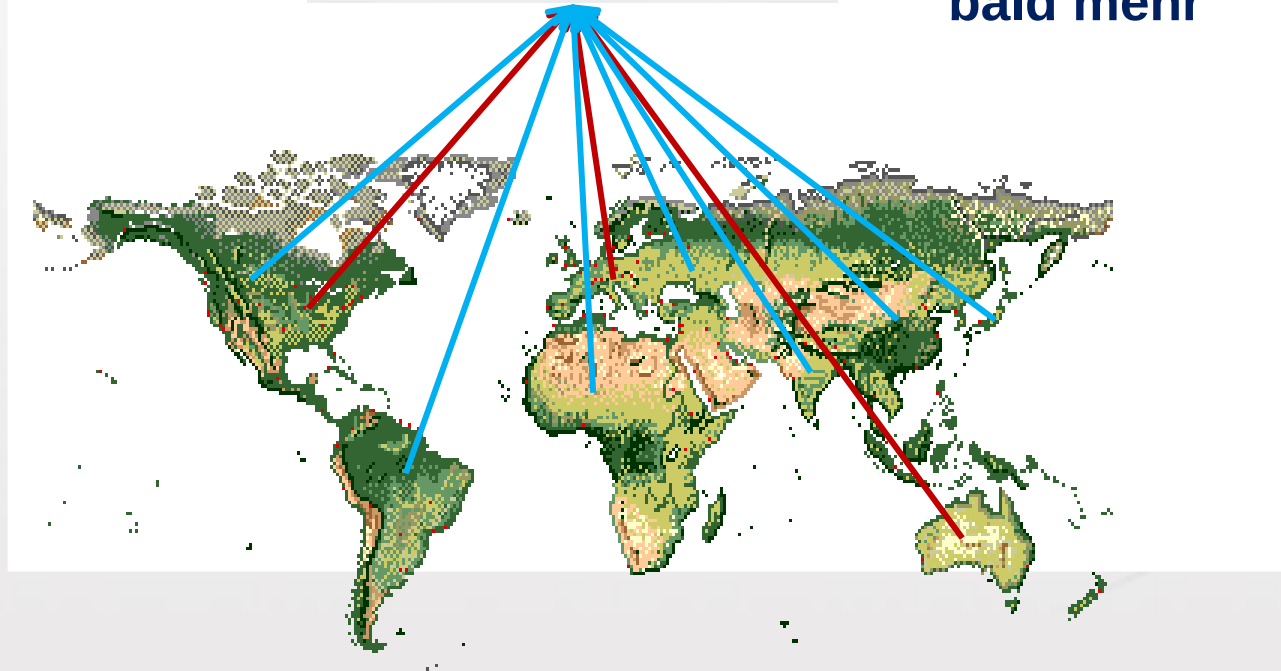


N. Kroes
VP EU Commission



**gegenwärtige
Förderer:**
**EC,
NSF/NIST,
AU ministry**

bald mehr



welcher Weg?

- welche Wege gibt es, um Hürden zu überwinden?
 - glauben wir an EIN top-down System – NEIN!
 - nur ein bottom-up Ansatz kann funktionieren
 - sehr gutes Beispiel: Internet
- RDA hat die Ambitionen, um Daten-Wissenschaftler & -Bibliothekare zusammen zu bringen und schrittweise Barrieren aus dem Weg zu räumen
- RDA muss global und cross-disziplinär sein, denn auch die Wissenschaften sind derart organisiert

was machen wir in RDA?

- Idee ist, dass Community und Technology Experten
 - **Working Groups** mit Ergebnissen nach 18 Monaten definieren
der Fokus ist auf konkreten Resultaten, die Hürden überwinden
 - **Interest Groups** definieren um Ideen auszutauschen die z.B. in konkrete WGs münden könnten
 - zu den **Plenaries** kommen und über Gruppen hinweg zu arbeiten
 - Zwischentreffen zu veranstalten, um Diskussionen voranzutreiben
 - Zwischen- und End-Resultate erzeugen
- wir müssen mit vielen interagieren
 - Wissenschaftliche Organisationen (3. Meeting geplant)
 - Wissenschaftlern (gerade Workshop mit 17 Top-Wissenschaftlern)
 - Industrie (ist in der Planung)
 - Policy Level (kontinuierlicher Austausch – aber auch Staten)
 - Teilnahme an vielen Community-Veranstaltungen

RDA Working Groups

- Data Foundation and Terminology
 - PID Information Types
 - Data Type Registries
 - Practical Policy
 - Metadata Standards Directory
 - Data Description Registry Interoperability
 - Standardisation of Data Categories and Codes
 - Data Citation
 - Wheat Data Interoperability
- Data Domain Terminology
 - Harmonized API
 - Flexible Registries
 - Recommended Policies
 - Overview about Metadata
 - finding data across multiple registry systems
 - ISO 639-6
 - Citation Recommendation
 - defining community standards

einige neue WGs haben sich bereits angemeldet

RDA Interest Groups

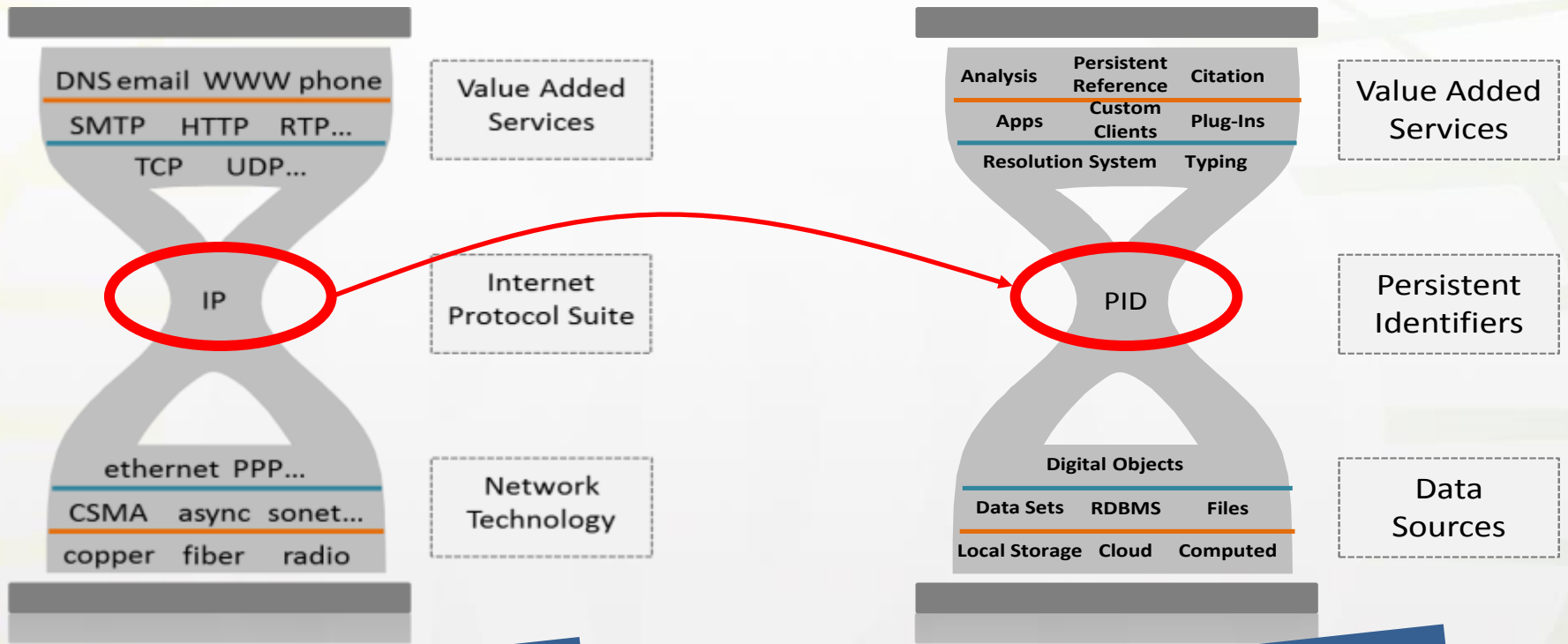
Cross-Domain IGs

- Research Data Provenance
- Big Data Analytics
- Brokering
- Data in Context
- Long tail of research data
- Metadata
- Preservation e-Infrastructure
- Semantic Interoperability
- PIDs
- Federated Identity Management
- Publishing Data
- Certification of Digital Repositories
- Community Capability Model
- Development of cloud computing capacity and education for developing world
- Engagement Group
- Legal Interoperability

Domain Specific IGs

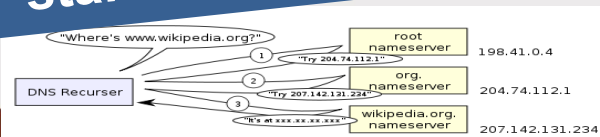
- Toxic genomics Interoperability
- Structural Biology
- Agricultural Data Interoperability
- Biodiversity Data Integration
- Defining Urban Data Exchange for Science
- Digital Practices in History and Ethnography
- Marine Data Harmonization

Lernen vom Internet - welche Ebene



Internet Domain
nodes with IP numbers
packages being exchanged
standardized protocols

Data Domain
objects with PID numbers
objects being exchanged
standardized protocols



point to
each other

metadata
attributes

würden alle PIDs verwenden ...

- wir könnten unsere Software auf die Verwendung von PIDs einrichten
- wir könnten ein API für alle Repositorien definieren, um Daten-Objekte anzufordern
- wir könnten mit den PIDs allgemeine Attribute assoziieren
 - zum Prüfen der Identität und Integrität
 - zum Finden der Metadaten, der Landing Page etc.
 - zum Identifizieren des Original-Repositories
 - zum Finden der Rechte Information
 - etc.
- es würde genau einer dieser vielen kleinen Schritte hin zu einer von der Software unterstützten Daten-Fabrik sein

wie weit ist RDA?

■ Plenary 1: March 18-20, 2013

■ at Gothenburg, Sweden

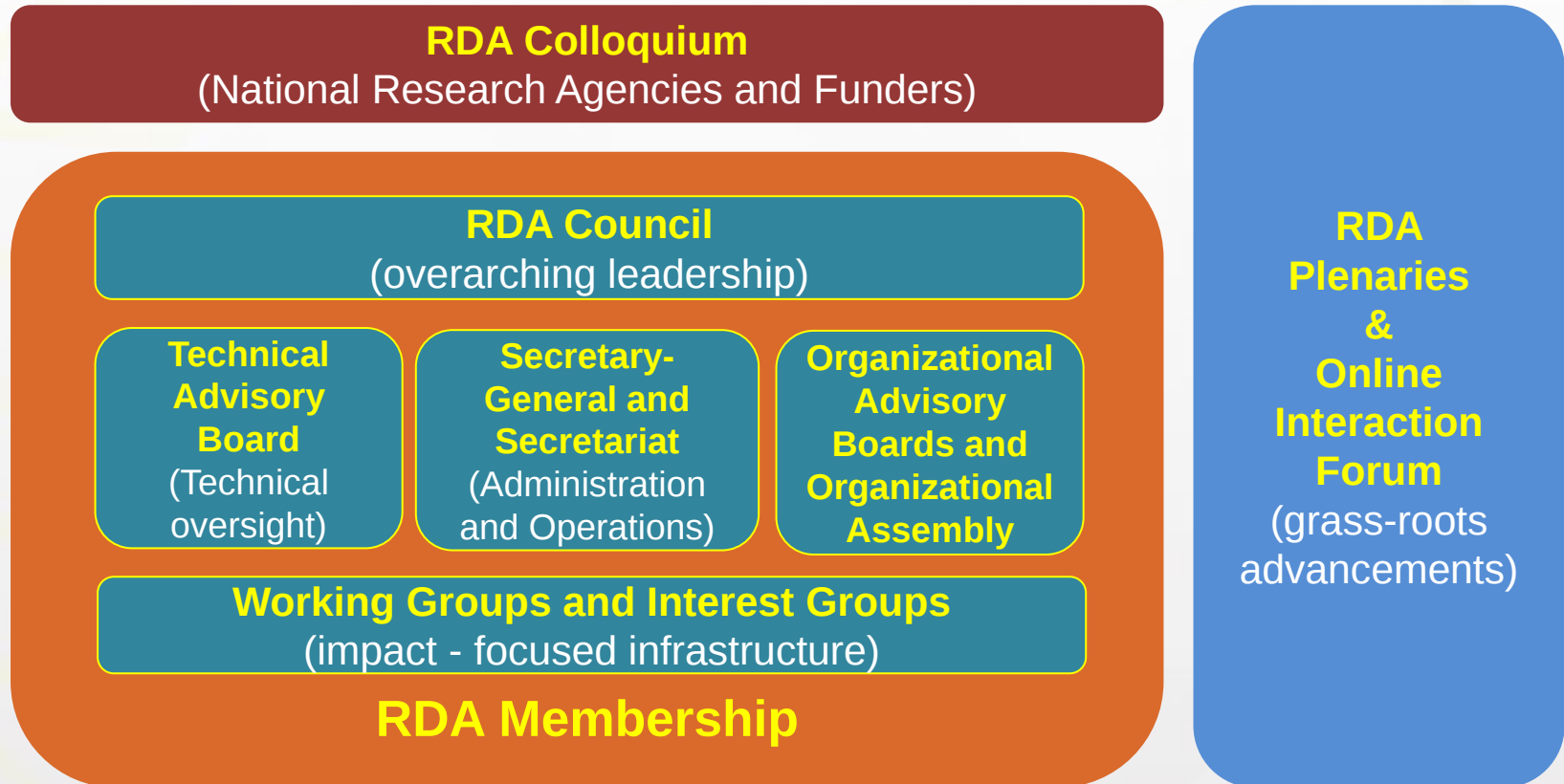
■ Plenary 2: September 16 - 18, 2013

■ at the National Academy of Sciences in Washington, DC, USA

■ Plenary 3: **Dublin, Ireland March 26-28, 2014**, hosted by RDA/Australia and Ireland

■ Plenary 4: **Amsterdam, September 22-24, 2014** hosted by RDA/Europe and NL erste Resultate zu erwarten

RDA Governance



alles ist aufgebaut – es gibt ziemlich viel Arbeit zu meistern

RDA Mitgliedschaft

❑ jeder der sich registriert und die Basis-Prinzipien (G8+O6) unterstützt, kann mitarbeiten

- **openness,**
- consensus-based decision making,
- balanced representation,
- technical neutrality,
- harmonization across communities and technologies,
- non-profit approach.

❑ Organisationen können Co-Funding Mitglieder werden

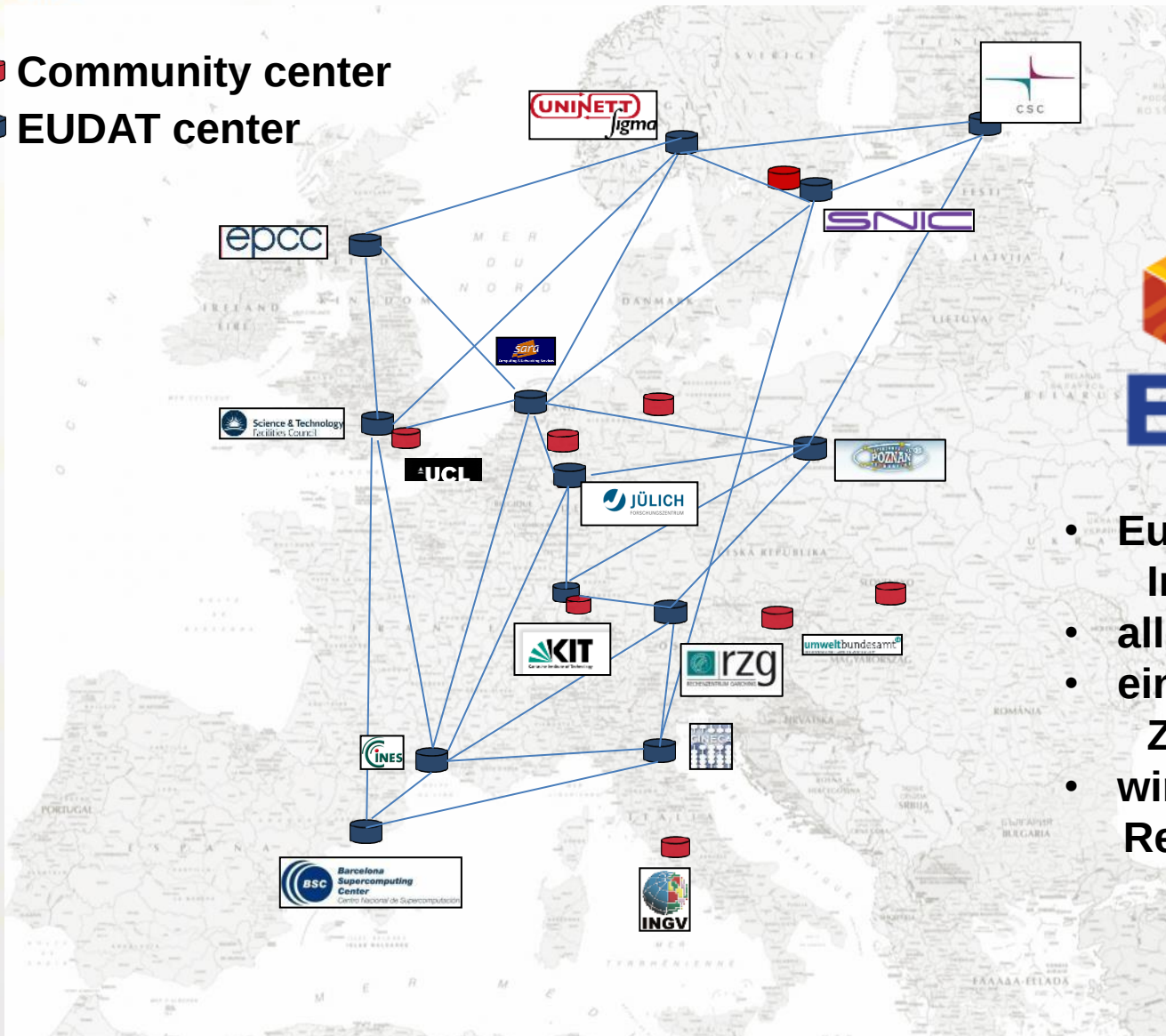
gibt es Risiken?

- ☐ RDA ist sehr jung – gerade mal 1+ Jahre
 - ☐ müssen es beschützen und bewässern
 - ☐ die RDA Pflanze kann sterben – klar!!!
-
- ☐ wir wissen nicht mal wie sie genau aussehen wird
 - ☐ aber wir müssen auch 5-10 Jahre voraus denken



und man braucht Infrastrukturen

-  Community center
-  EUDAT center



- Europäische Daten-Infrastruktur
- alle großen EU Zentren
- einige Community Zentren
- wir müssen RDA Resultate umsetzen

machen Sie mit?

- ☐ wenn Sie mit den Zielen von RDA übereinstimmen, dann
 - ☐ machen Sie auf RDA in Ihrer Community aufmerksam
 - ☐ fordern Sie ihre jungen Leute auf, teilzunehmen
 - ☐ machen Sie PR für Training Kurse etc.
- ☐ wenn Sie nicht übereinstimmen, dann
 - ☐ brauchen wir Ihre Kommentare und Rat
- ☐ es ist ja toll, dass sich der RatSWD einmischen will
- ☐ in der Tat müssen wir uns auch intensiv um die deutsche Forschungslandschaft kümmern

Dank für Ihre Aufmerksamkeit.

- Information: <http://rd-alliance.org>
- Fragen, Kommentare: enquiries@rd-alliance.org
- RDAAEurope-info@postit.csc.fi <http://europe.rd-alliance.org>

RDA Maschinerie

support

Support process

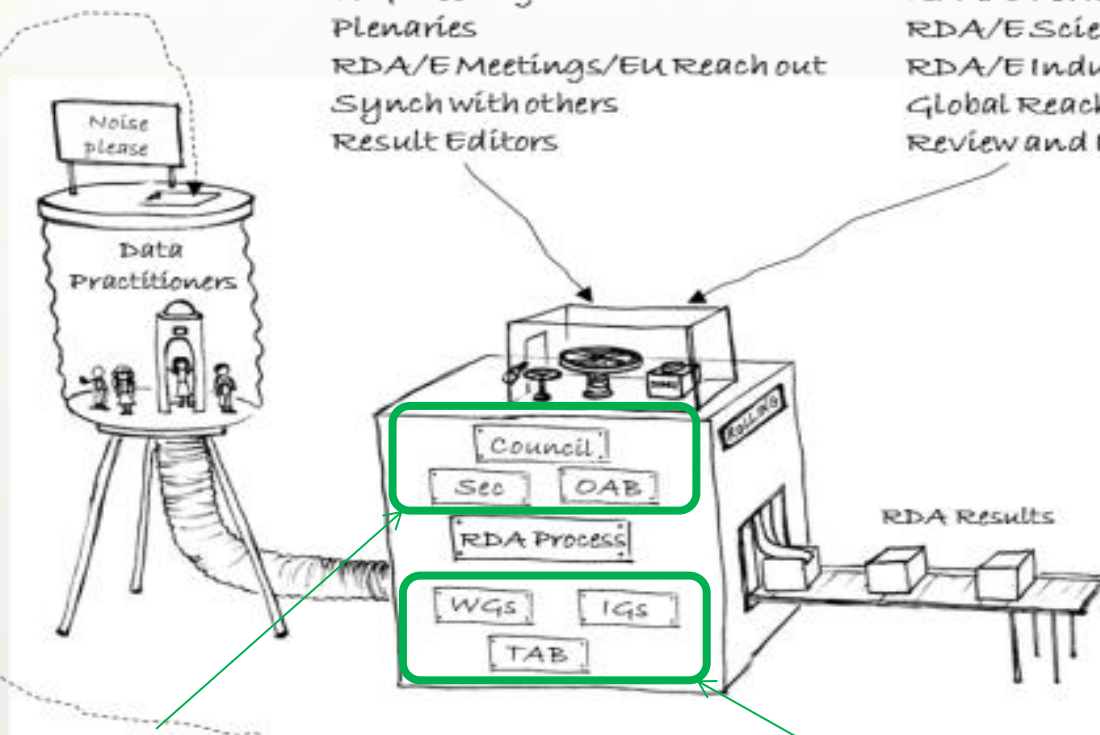
Council, OAB, TAB, Sec, Chairs
WG meetings
Plenaries
RDA/E Meetings/EU Reach out
Synch with others
Result Editors

Comment process

RDA/E Forum
RDA/E Science Workshop
RDA/E Industry Forum
Global Reach out
Review and Impact Summit

external
engagement

results
to come



Result Dissemination

Storing/Publishing Results
Offering Results
Gather Impacts
Training about results
Evangelization and Interaction
Help to implement results
Help to adapt to results
Help to adapt software

top-down
balance

bottom-up
work

The RDA-RDA/E Machine

Phase II: How to push the bottom-up work and uptake

RDA Resultate

- wir erwarten erste Resultate im September
- Art der Resultate und IP issues
 - **Discussion Documents** – CC-BY
 - RDA Policies, Working Group Case Statements, and Interest Group Charters – default: CC-BY
 - RDA **Recommendations** – approval required; CC-BY or CC0 or dual licensing?
 - **Implementations** – Open source preferably BSD Type