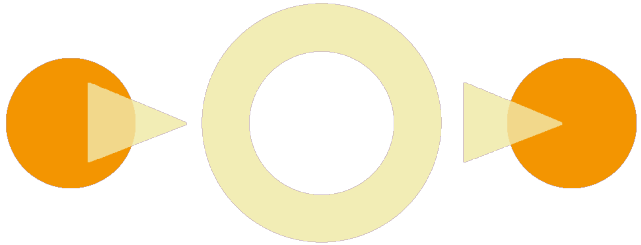
Sichtweisen von EPIC

RDA-DINI-Workshop



Ulrich Schwardmann

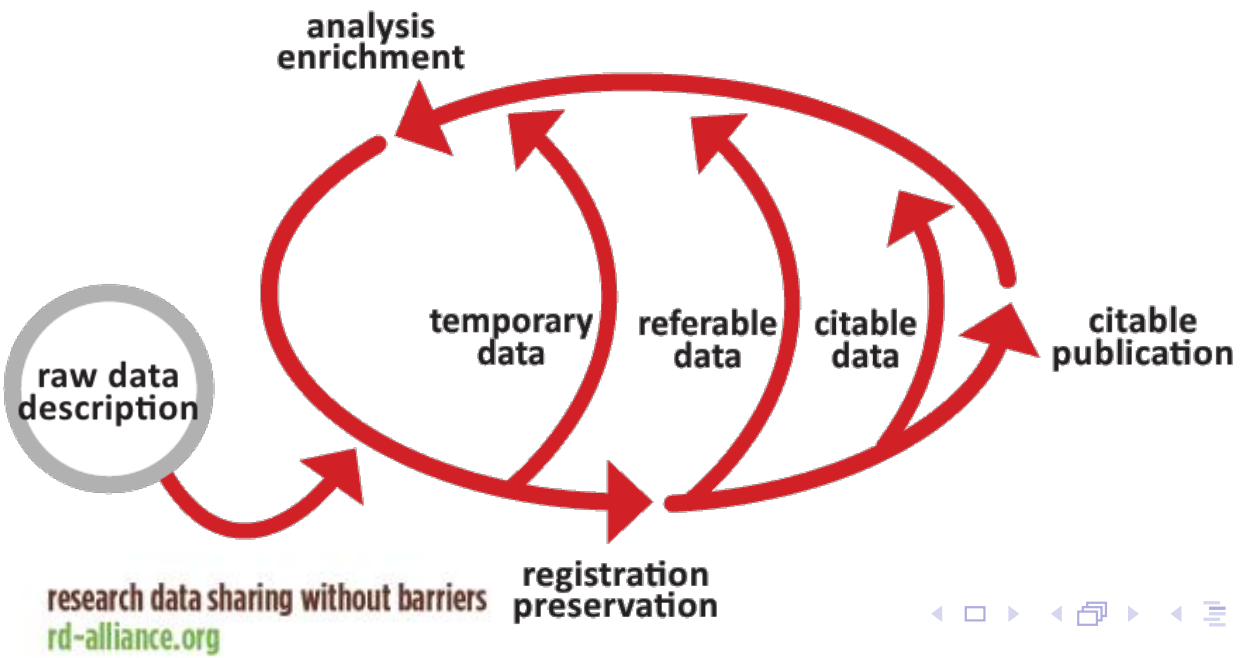
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

Am Fassberg, 37077 Göttingen
ulrich.schwardmann [at] gwdg.de

29 May 2015, Karlsruhe

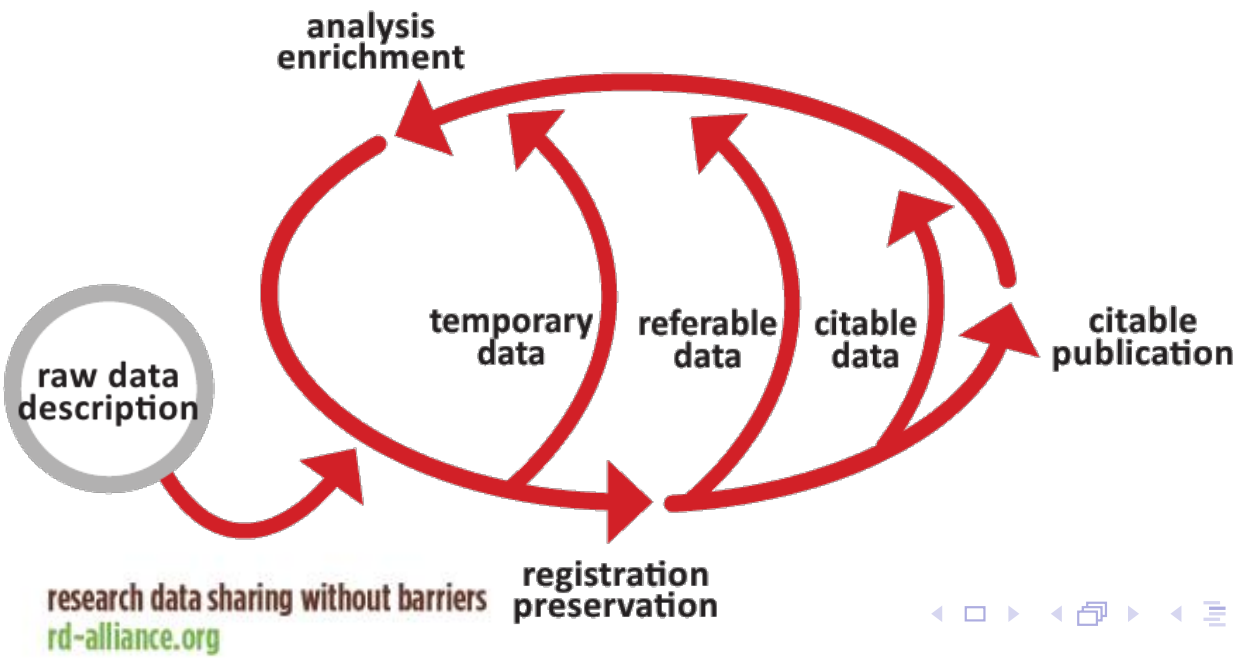
EPIC – Konsortium für eine PID Infrastrukture für eResearch

- Fokus: wissenschaftliche und kulturelle Comunities
- besteht derzeit aus sechs Europäischen Rechenzentren: *CSC (Finland), DKRZ (Germany), GR-Net (Greece), GWDG (Germany), PDC (Sweden), SURF-SARA (Netherlands)*
- liefert PIDs für das Datenmanagement digitaler Objekte
  - typischerweise sehr früh im wissenschaftlichen Prozess



EPIC – Konsortium für eine PID Infrastrukture für eResearch

- Fokus: wissenschaftliche und kulturelle Comunities
- besteht derzeit aus sechs Europäischen Rechenzentren: *CSC (Finland), DKRZ (Germany), GR-Net (Greece), GWDG (Germany), PDC (Sweden), SURF-SARA (Netherlands)*
- liefert PIDs für das Datenmanagement digitaler Objekte
  - typischerweise sehr früh im wissenschaftlichen Prozess



EPIC Infrastructure

- EPIC gibt selbst Prefixes aus im Rahmen von DONA durch GWDG als MPA
- EPIC PIDs haben eine hohe Zuverlässigkeit
  - durch ein langfristiges Commitment der Partner
  - durch Spiegelung der local handle server (LHS) (in der Regel zweimal)
  - durch Spiegelung der global handle registry (GHR)
  - und des handle resolution proxy (beide bei GWDG, die ersten ausserhalb USA)
    - beides verbesserte die mittlere Auflösungszeit innerhalb Europas drastisch
  - durch Audits
- **EPIC** arbeitet eng mit **EUDAT** and **RDA** zusammen

EPIC Infrastructure

- EPIC gibt selbst Prefixes aus im Rahmen von DONA durch GWDG als MPA
- EPIC PIDs haben eine hohe Zuverlässigkeit
  - durch ein langfristiges Commitment der Partner
  - durch Spiegelung der local handle server (LHS) (in der Regel zweimal)
  - durch Spiegelung der global handle registry (GHR)
  - und des handle resolution proxy (beide bei GWDG, die ersten ausserhalb USA)
    - beides verbesserte die mittlere Auflösungszeit innerhalb Europas drastisch
  - durch Audits
- **EPIC** arbeitet eng mit **EUDAT** and **RDA** zusammen

Zentrale Rolle der PIDs

Scholarly Context Not Found[1]

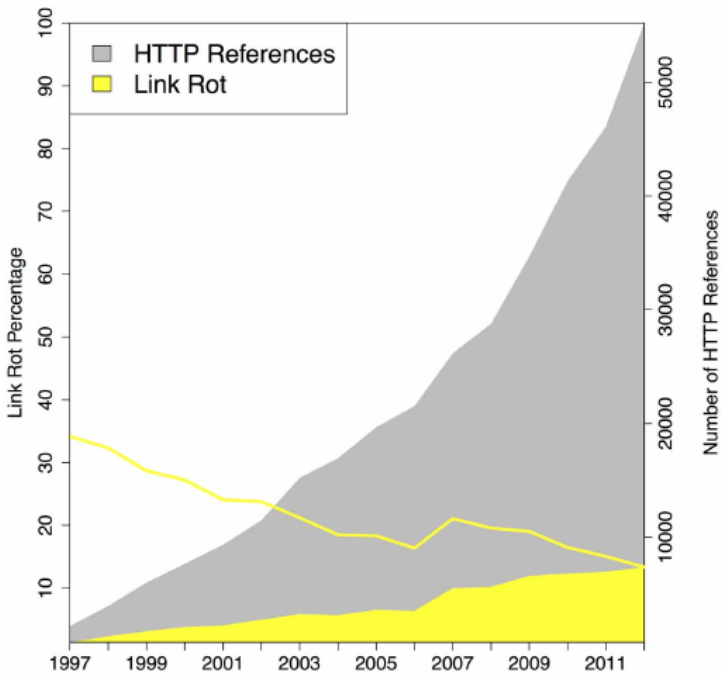


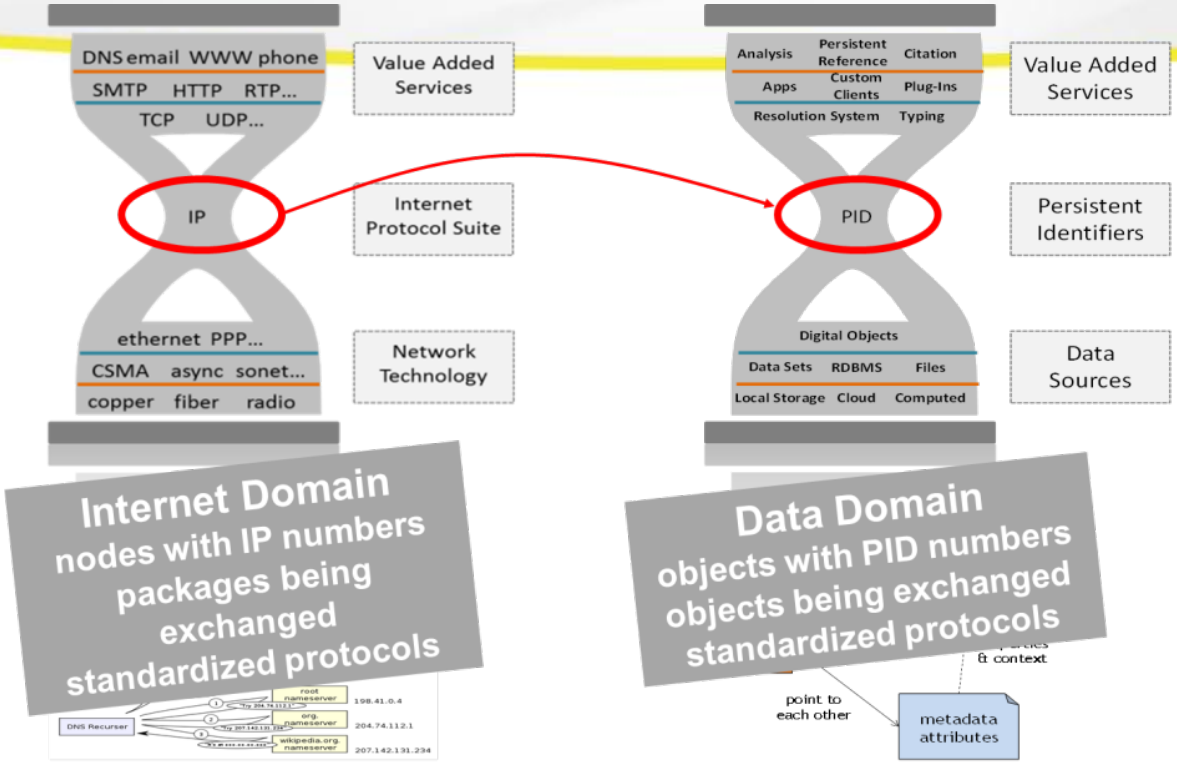
Fig. 10. Link Rot - arXiv corpus.

doi:10.1371/journal.pone.0115253.g010

- untersucht **link rot** und **content drift**
  - bei drei Corpora aus Science, Technology, and Medicine (STM)
  - zwischen den Publikationsjahren 1997 bis 2012
  - mit 3.5 Mio Artikeln, davon 1 Mio mit Referenz zu Web-Resourcen
- 7 von 10 STM Artikel mit Web-Referenz leiden unter Reference Rot (1 von 5 insgesamt)

[1] Klein,Van De Sompel,ea.: *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot (2014)*, DOI10.1371/journal.pone.0115253

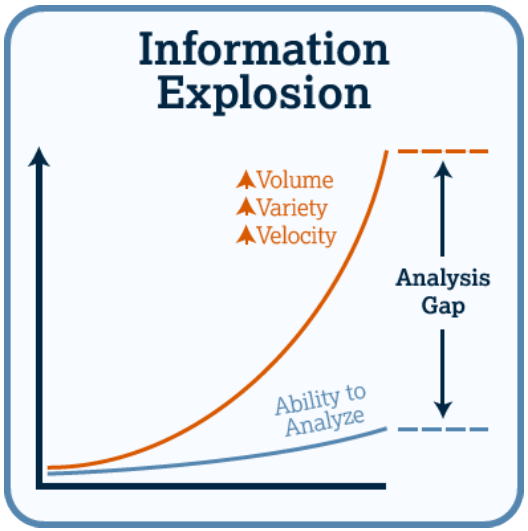
Zentrale Rolle der PIDs



- Die Analogie:
  - IP-Adressen erlauben eine Vielzahl an Netzwerkdiensten, verfügbar über viele verschiedene Netzwerke
  - PIDs erlauben eine Vielzahl an Datenmanagement-Diensten, verfügbar über viele verschiedene Datenquellen

Die Informations Explosion: The Gap

RDA:



- ... the **increasing gap** between the amount of **data that we produce** and the amount of data that we are actually **capable of analysing** ... require(s) **new strategies** for managing data if we are **to keep up with what we generate** ... [a]

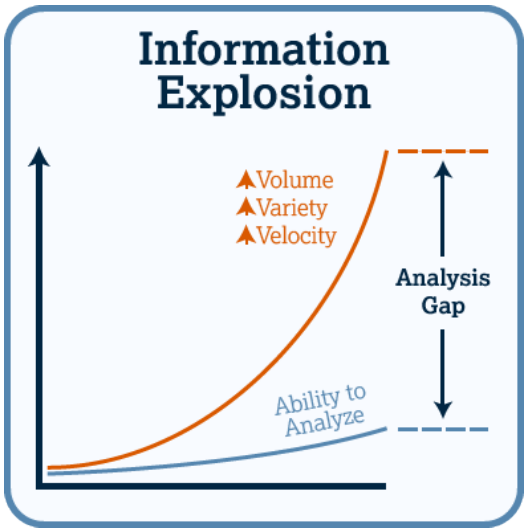
[a] RDA-DFIG White paper

- ... currently the data growth is higher than Kryder's (Moore's) Law ... (Margeret Leinen at RDA P5)
- implizites Ziel von RDA ist **die Lücke zu schliessen**

EPIC zu RDA
Ulrich Schwardmann
EPIC
EPIC – Konsortium
Zentrale Rolle der PIDs
Informations Explosion
The Gap
Entwicklungen beim Speicher
Cloud/Objekt-Speicher
Manage the Gap
Diskussion

Die Informations Explosion: The Gap

RDA:



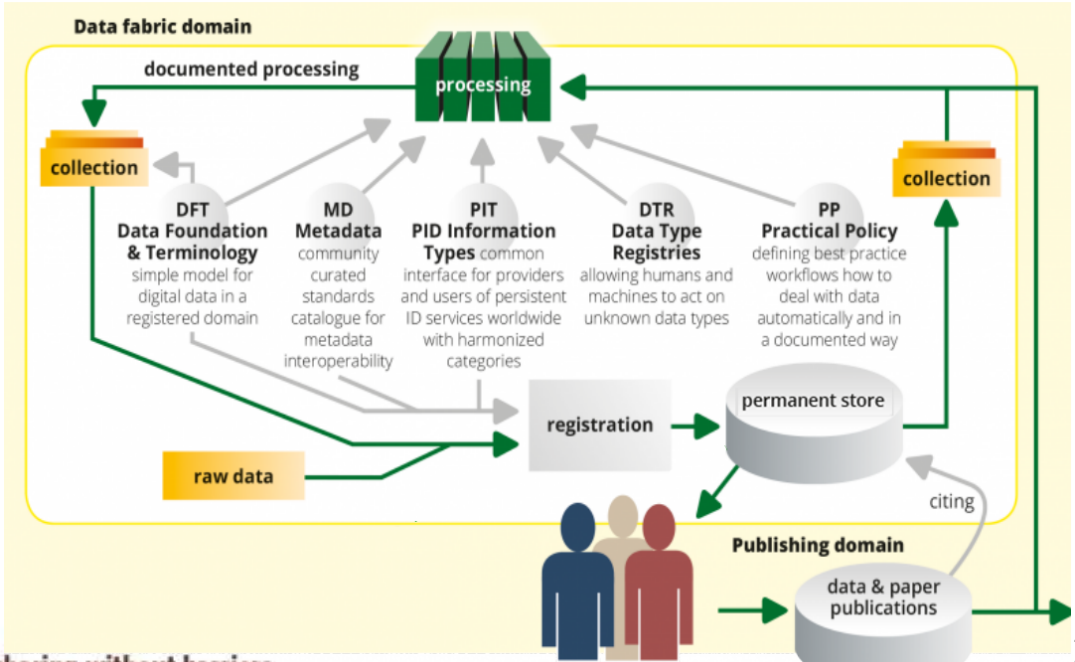
- ... the **increasing gap** between the amount of **data that we produce** and the amount of data that we are actually **capable of analysing** ... require(s) **new strategies** for managing data if we are **to keep up with what we generate** ... [a]

[a] RDA-DFIG White paper

- ... currently the data growth is higher than Kryder's (Moore's) Law ... (Margeret Leinen at RDA P5)
- implizites Ziel von RDA ist **die Lücke zu schliessen**

RDAs Schritte, die Lücke zu schliessen

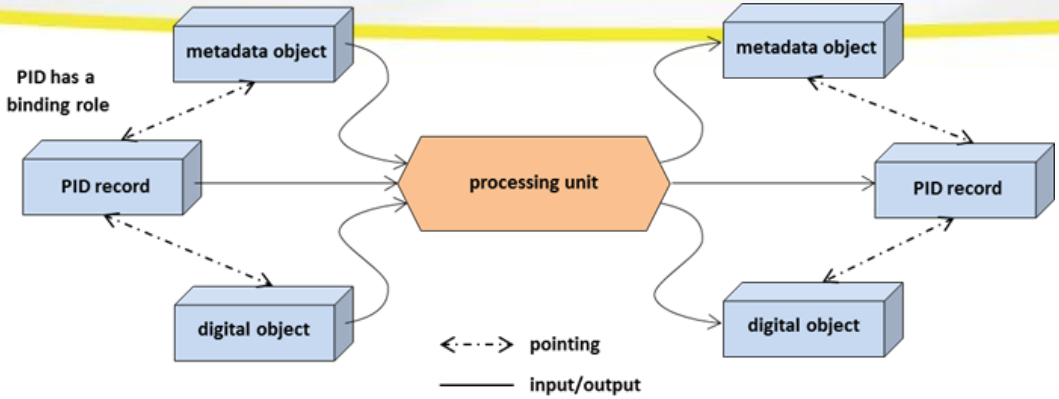
- Die größte Diskrepanz ist in der Datenanalyse die **Unauffindbarkeit** und **mangelnde Nachnutzbarkeit**
  - *reference rot* und *content drift*
  - Metadaten sind oft nicht vorhanden, nicht interpretierbar oder nicht mit den Daten verbunden
- **Registrierung**: Verbindung von Metadaten und Daten mit einem PID zu einem DO
- halte registrierte Daten dauerhaft interpretierbar



[2]

[2] RDA-DFIG White paper

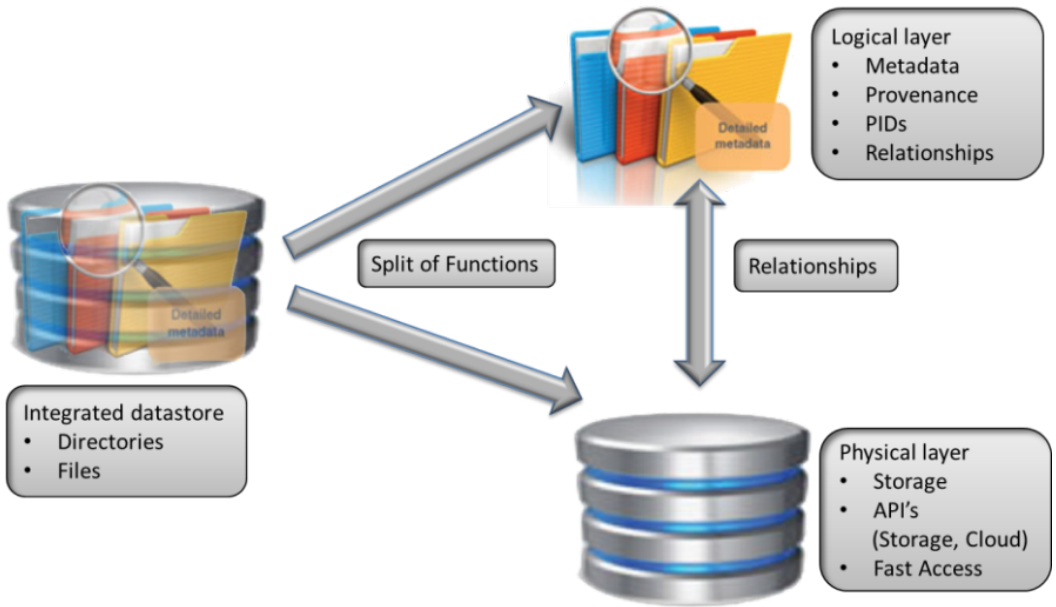
Datenkurations-Workflows



- Das *Digital Object* ist das DO-PID-MD Triple
- PID sind in der Mitte, weil
  - der Zeiger zum DOI im Workflow benutzt wird
- Die Parametrizierung des Workflows
  - geschieht über stark standardisierte Metadaten
  - die sehr nah am Pointer sein sollten
  - verwende PID information types (PIT) mit registrierten Types
- EPIC erlaubt PIT und Referenz zu Daten wie Metadaten

Entwicklungen bei der Speichertechnologie
Cloud, Objekt-Speicher

- spaltet Datenhaltung in physikalischen und logischen Layer



- vereinfacht und optimiert lokalen Speicherzugriff
- hilft bei Nachnutzbarkeit: verbindet Daten und Metadaten
- hilft teilweise bei Registration, weil Metadaten generisch eingebunden sind

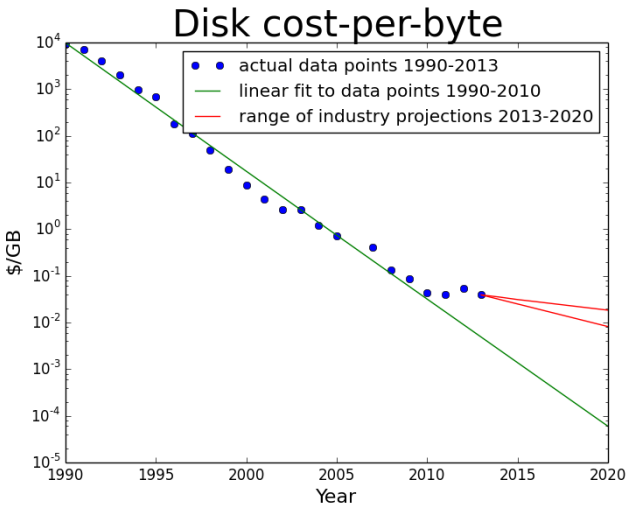
Entwicklungen bei der Speichertechnologie
Cloud, Objekt-Speicher

- nutzt eine interne ID (hash) ohne globale Auflösung
- nutzt *flüchtige* URLs für globale Adressierung
- nutzt weniger bei der Auffindbarkeit (reference rot durch flüchtige URLs)
- BTW: viele Repository-Systeme haben den gleichen Ansatz
- Eine Konzept für eine Verbindung zwischen interner ID und globaler PID ist nötig
  - am besten durch ein Interface zwischen ID-Datenbank und PID System
  - und wünschenswert: Implementierung von (P)ID Information Types
- Notwendig für RDA: eine engere Einbindung von Speicher-Herstellern, Repository-System-Herstellerns, und Kooperation mit PID-Providern und Datenzentren

Entwicklungen bei der Speichertechnologie
Cloud, Objekt-Speicher

- nutzt eine interne ID (hash) ohne globale Auflösung
- nutzt *flüchtige* URLs für globale Adressierung
- nutzt weniger bei der Auffindbarkeit (reference rot durch flüchtige URLs)
- BTW: viele Repository-Systeme haben den gleichen Ansatz
- Eine Konzept für eine Verbindung zwischen interner ID und globaler PID ist nötig
  - am besten durch ein Interface zwischen ID-Datenbank und PID System
  - und wünschenswert: Implementierung von (P)ID Information Types
- Notwendig für RDA: eine engere Einbindung von Speicher-Herstellern, Repository-System-Herstellerns, und Kooperation mit PID-Providern und Datenzentren

Entwicklungen bei der Speichertechnologie



- Kryder's Law scheint seit 2011 nicht mehr gültig zu sein
- die Hoffnung: weniger Daten könnten eine kleinere Lücke bedeuten
- Aber zur Erinnerung: die Lücke entstand zwischen Produktion (nicht Speicher) und Analyse

- Aussicht:
  - die Lücke wird zusätzlich größer, und wird sich wohl nicht schliessen lassen[3]
  - Frage: wie kann der **Datenverlust optimal gesteuert** werden?
- RDA sollte auch hierfür neue Wege aufzeigen

[3] David Rosenthal, blog.dshr.org

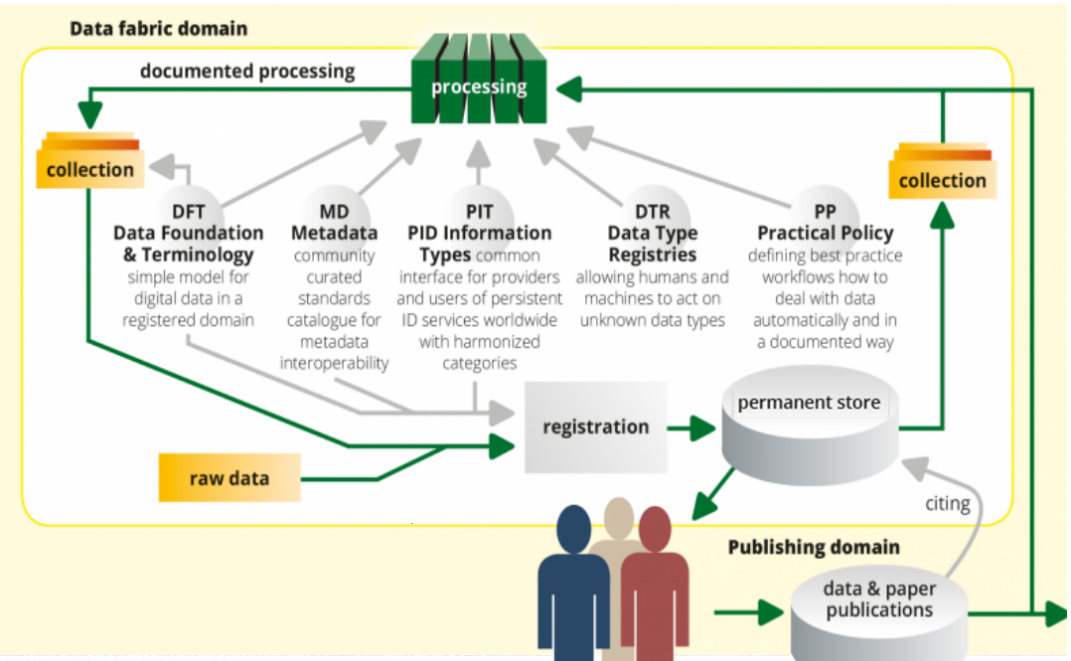
Manage the Gap

Thesen:

- es wird nicht möglich sein, alle wissenschaftlichen Daten dauerhaft zu speichern
- ihr wissenschaftlicher Wert ist nicht a priori bekannt
  - und kann mit der Zeit schwinden.
- Beispiele
  - teure, aber (beinahe) vollständig reproduzierbare Prozesse
    - Resultate von High End Simulationen
    - NGS
  - Daten, die bei der Findung von Ergebnissen eine Rolle gespielt haben, aber nicht mehr bei der Überprüfung?
  - Daten mit geringer Signifikanz
  - ...
- ausgenommen: Datensätze, auf die sich Veröffentlichungen beziehen

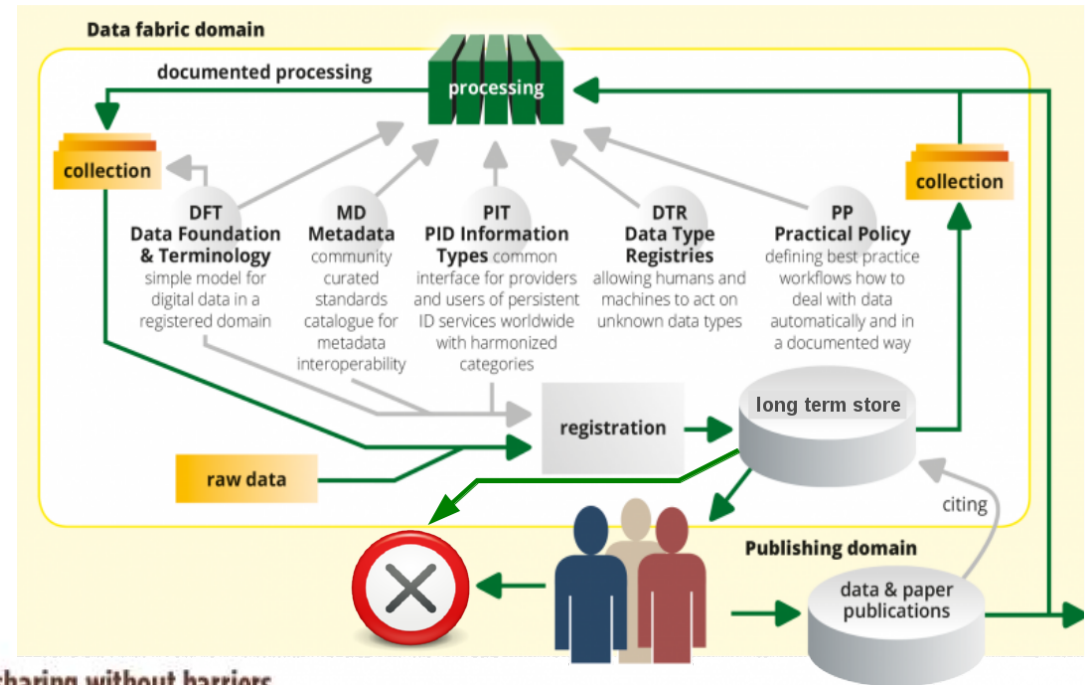
Manage the Gap

- ein transparenter Auswahlprozess ist nötig
  - dazu wird vor allem eine öffentlich bekannte Lebenszeit (expiration date) der Daten benötigt
  - Entscheidungen über die Lebenszeit müssen durch die gesamte wiss. Community beeinflussbar sein (Reviewing)
- auch nach der Löschung sollte es noch Hinweise (Metadaten) über die Daten geben (Tombstones)
- PIDs sind auch hier das entscheidende Instrument



Manage the Gap

- ein transparenter Auswahlprozess ist nötig
  - dazu wird vor allem eine öffentlich bekannte Lebenszeit (expiration date) der Daten benötigt
  - Entscheidungen über die Lebenszeit müssen durch die gesamte wiss. Community beeinflussbar sein (Reviewing)
- auch nach der Löschung sollte es noch Hinweise (Metadaten) über die Daten geben (Tombstones)
- PIDs sind auch hier das entscheidende Instrument



Zusammenfassung

- wichtigste RDA Outcomes für EPIC
  - Data Type Registry, PID Information Types, Policies
- neue Komponenten und Verantwortlichkeiten:
  - DTR: Aufbau eines Produktivbetriebs (EPIC)
  - PID Info Types: Aufbau eines Produktivbetriebs
    - Schnittstellen (EPIC)
    - Standards (Community-Process)
  - Ableitung von Policies
    - Schnittstellen (EPIC)
    - Standards (Community-Process)
    - Beispiel: Kontrollierte Datenlöschung
  - PID in der Speichertechnologie (Object Storage und NDN)
- neue benötigte RDA-Aktivitäten
  - Speichertechnologie, Kurationstechnologie, PIT-parametrisierte Workflows (Policies)
- Kollaborationsmöglichkeiten
  - Communities, Speicherhersteller, DataCite

Fragen und Diskussion

Vielen Dank für Ihre Aufmerksamkeit

Hinweis auf eine geplante, gemeinsame

PID-Konferenz von EPIC und DataCite

am 21.9.2015 in Paris (kurz vor RDA P6)