kobv Kooperativer Bibliotheksverbund Berlin-Brandenburg

ZIB

Ist auch drin was draufsteht? Qualitätssicherung bei der Langzeitarchivierung von Forschungsdaten

Wolfgang Peters-Kottig, Tim Hasler

Kooperativer Bibliotheksverbund Berlin Brandenburg
am Zuse-Institut Berlin

Bibliothekskongress Leipzig 2013

Hintergrund: Projekt EWIG

»**E**ntwicklung von **W**orkflow-komponenten für die Langzeit-archivierung von Forschungsdaten **i**n den **G**eowissenschaften«

Projektpartner:

- GeoForschungsZentrum Potsdam
- Institut f. Meteorologie der FU Berlin
- KOBV

Foto: Matthias Groß

Hintergrund: Projekt EWIG

- Entwicklung von institutionellen Policies
- Konzepte für Lehrveranstaltungen zum Forschungsdatenmanagement in Fachdisziplinen
- Qualitätssicherung beim Ingest von Forschungs-daten

Erkenntnisse aus Expertengesprächen

- Unterschiede zwischen Wissenschaftsdisziplinen:
  - »Verwendbare Dateiformate«, Aufwand bei Metadaten-beschreibung, Umfang der Datenkuratierung
    - → Fachstandards (Fachkultur) bei der Übergabe von Forschungsdaten
  - Stand der Langzeitarchivierung in einzelnen Fächern gut, in den meisten weiter in »experimenteller Phase«

Erkenntnisse aus Expertengesprächen

- Wunsch der Wissenschaftler nach Checklisten mit konkreten Handlungsempfehlungen
- Fächerübergreifender Austausch von Daten und technische Interoperabilität sind »ausbaufähig«
- Bei Gesprächen über Daten fehlt eine gemeinsame Sprache
- Technische Qualitätssicherung bei Datenübergabe wird als Aufgabe auch des Datenproduzenten wahrgenommen

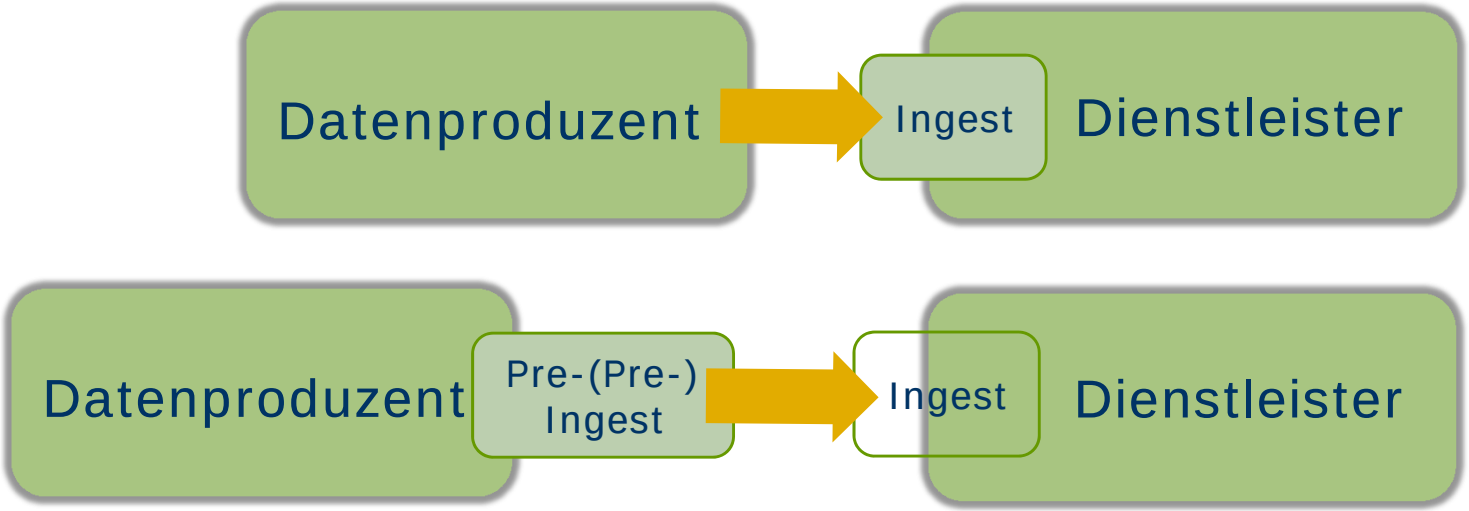
Erkenntnisse aus Expertengesprächen

Konzept des Datenmanagement mit Dateipaketen (Objektbezogene Verarbeitung) greift häufig nicht

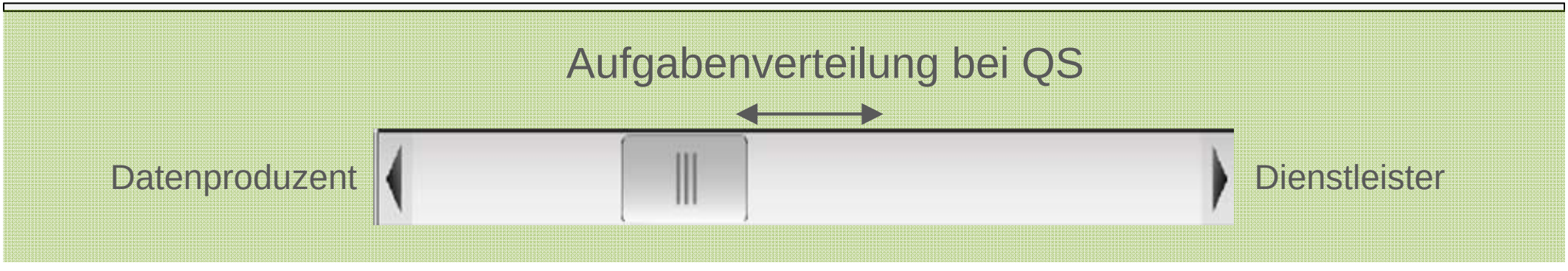
0001a	Roaring Creek	USA	Illinois Basin	Brazil Fm.	Atokan, U.	311.3	311.0	-24.4	coal	tropical
0001b	Roaring Creek	USA	Illinois Basin	Brazil Fm.	Atokan, U.	311.3	310.4	-24.6	cuticles	tropical
0001c	Roaring Creek	USA	Illinois Basin	Brazil Fm.	Atokan, U.	311.3	310.4	-24.3	coalified tissue	tropical
0001d	Roaring Creek	USA	Illinois Basin	Brazil Fm.	Atokan, U.	311.3	310.4	-24.6	cuticles	tropical
0001e	Roaring Creek	USA	Illinois Basin	Brazil Fm.	Atokan, U.	311.3	310.4	-24.6	cuticles	tropical
0001f	Roaring Creek	USA	Illinois Basin	Brazil Fm.	Atokan, U.	311.3	310.4	-25.1	cuticles	tropical
2	Borehole; Vilui River	Russia			Tatarian, L.	262.8	261.3	-25.8	coalified tissue	cool temperate
3	Borehole PK-640; Usa River	Russia	Pechora Basin		Kazanian	266.5	268.8	-22.8	cuticles	cool temperate
0004a	Lone Star Lake Spillway	USA	Illinois Basin	Lawrence shale Fm.; Douglas Gp.	Virgilian	299.0	301.9	-24.2	cuticles	tropical
0004b	Lone Star Lake Spillway	USA	Illinois Basin	Lawrence shale Fm.; Douglas Gp.	Virgilian	299.0	301.9	-24.1	cuticles	tropical
7		Russia	Moscow Basin	Tula Group	Visean, U.	332.0	329.4	-24.6	coal	tropical
8		Russia	Moscow Basin	Tula Group	Visean, U.	332.0	329.4	-23.0	coal	tropical
9		Russia	Moscow Basin	Tula Group	Visean, U.	332.0	329.4	-22.0	coal	tropical
10	Workuta	Russia	Pechora Basin		Kazanian	266.5	268.8	-23.0	coal	cool temperate
11	Dickson Land	Norway		Mumien Fm.; Andrée Land Gp.	Visean, U.	332.0	329.4	-22.3	coal	tropical
13	Borna-Hainichen	Germany	Erzgebirge Basin	Hainichen Subgp.	Visean, U.	330.7	328.0	-22.9	coal	tropical
15	Butterloch/Pletterbach	Italy			Guadalupian-	260.8	260.6	-22.6	cuticles	arid

Erkenntnisse aus Expertengesprächen

Tendenz zu Vorverlagerung der tatsächlichen Datenübergabe in Richtung Produzent (=Wissenschaftler)



Qualitätssicherung bei der Datenübergabe



Inhaltliche QS

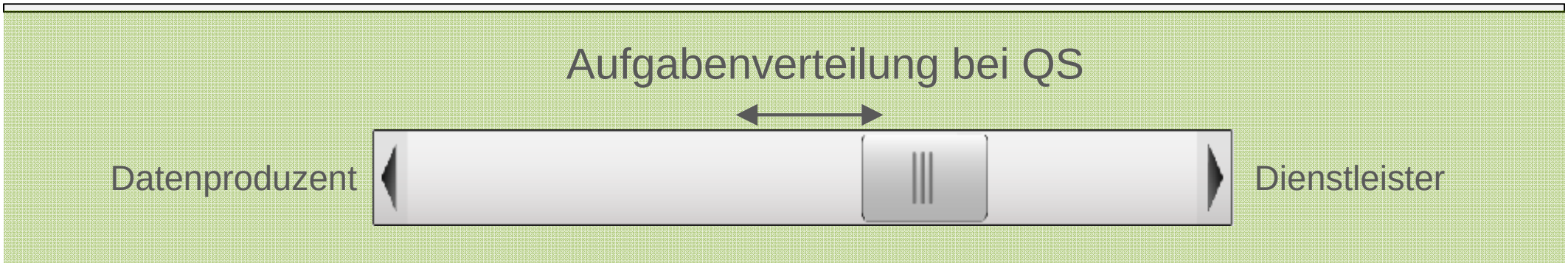
- Prüfung auf inhaltliche Richtigkeit
- Prüfung auf Datenplausibilität
  Erkennen v. Ausreißern und physikalisch unmöglichen Werten
  …Formatierungsfehler

ELEVATION 2: EXPEDITION: CON01-6 BASIS: R/V Vereshchagin GEA

Activity	Elevation	Elevation Bas	Magnetic susceptibility
	m	m	SI*10^-6
CON01-603-2	1.005		239.362
CON01-603-2	1.005		235.357
CON01-603-2	1.006		239.608
CON01-603-2	1.006		237.461
CON01-603-2	1.007		246.604
CON01-603-2	1.007		237.312
CON01-603-2	1.008		258.1
CON01-603-2	1.008		239.412
CON01-603-2	1.009		269.596
CON01-603-2	1.009		243.763
CON01-603-2	01. Jan		276.592
CON01-603-2	01. Jan		250.364
CON01-603-2	1.011		283.588
CON01-603-2	1.011		259.215
CON01-603-2	1.012		292.839

kobv Kooperativer Bibliotheksverbund Berlin-Brandenburg

Qualitätssicherung bei der Datenübergabe



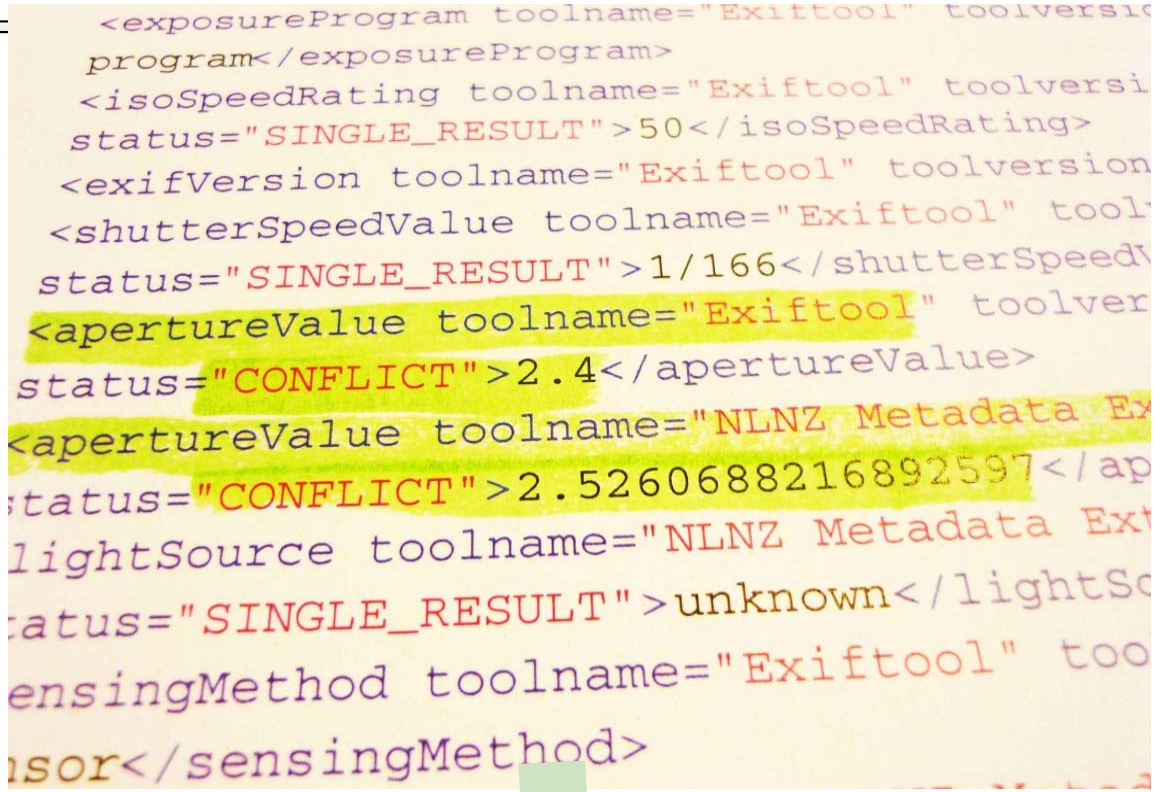
Technische QS

- Prüfsummen, Virencheck
- Formatcharakterisierung und Validierung von Dateien
- Metadatenextraktion
- Metadatenanreicherung

Tools für technische QS: Beispiel FITS

»File Information Tool Set« konfigurierbar, nutzt z.B.

- DROID
- JHOVE
- NLNZ Metadata Extractor
- ExifTool
- File utility (windows)
- FFIdent



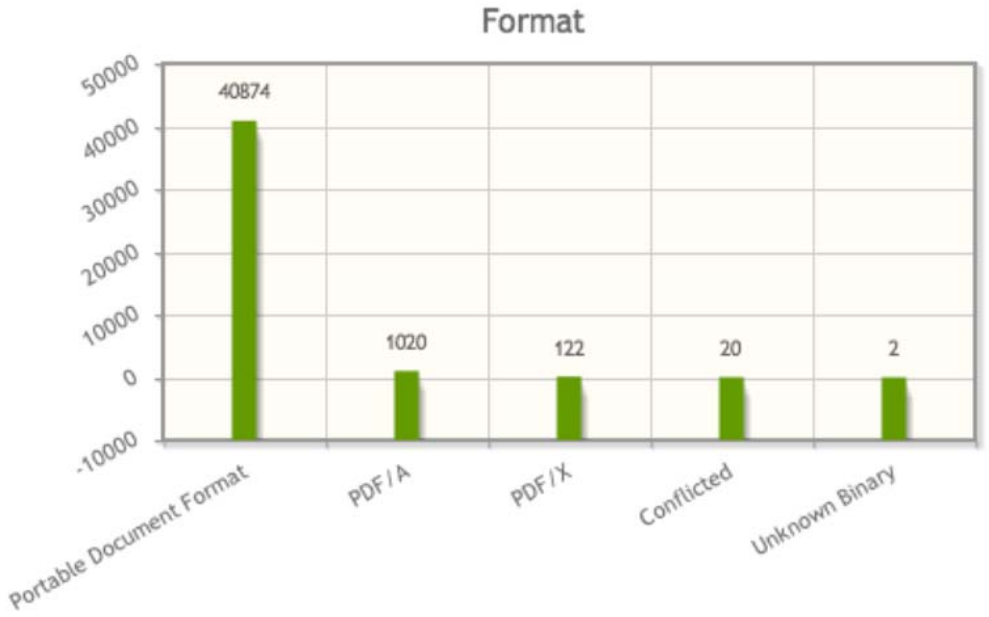
Was tun??

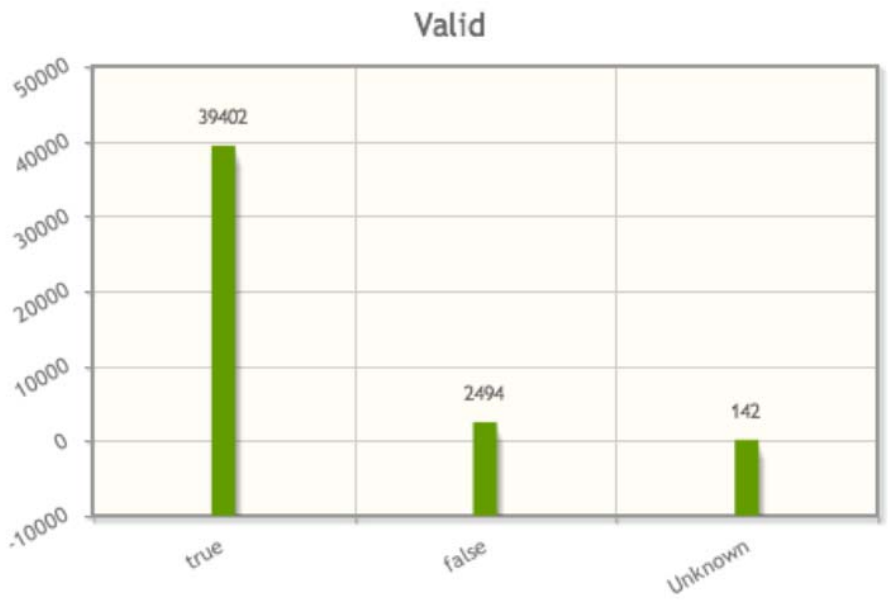
kobv Kooperativer Bibliotheksverbund Berlin-Brandenburg

ZIB

Tools für technische QS: Beispiel c3po

»Clever, Crafty Content Profiling of Objects« (http://ifs.tuwien.ac.at/imp/c3po)
Infos über Inhalt einer Dateisammlung
verwendet FITS-Output (auch andere möglich)





kobv Kooperativer Bibliotheksverbund Berlin-Brandenburg

ZIB

Probleme beim Einsatz von QS-Tools

- Fehlende Auswahlkriterien: Welche Tools einsetzen?
- Unvollständigkeit und geringe Usability
- Geringe Performanz, Präzision und Robustheit
- Unverständliche, überkomplexe Rückmeldungen
- Keine Empfehlungen für weitere Schritte

Wie weiter?

- (Weiter-)entwicklung von Toolwrappern, Webservices…
- technische QS als Service anbieten (Aufgaben vor-verlagern an den Datenproduzenten)
- Unterstützung für weitere Datei-Formate (z.B. Videos, komplexe Dateiformate)
- Empfehlungen, Handreichungen, Checklisten anbieten

…Vielen Dank!

http://ewig.gfz-potsdam.de
ewig-projekt@zib.de

Datenportale – Segen oder Fluch? Ein Bewertungsschema zur Nutzerfreundlichkeit.

Posted on 05.02.2013 by petra.gebauer

Im Zuge der Diskussionen um Open Access – Pro und Kontra – wird neben dem freien Zugang zu wissenschaftlicher Literatur auch die Bereitstellung wissenschaftlicher Daten (Mess- und Modelldaten) ergänzt durch Metadaten im Internet gefordert (Berliner Erklärung 2003).

Gerade in den Geowissenschaften werden zunehmend Daten über sogenannte Datenportale zur Verfügung gestellt. Sie unterscheiden sich natürlich bezüglich Art und Umfang der angebotenen Daten. Aber auch die Nutzbarkeit dieser Datenportale ist sehr unterschiedlich. So werden zwar vermehrt Datensätze über das Internet verfügbar, der Vorteil gegenüber dem direkten Datenaustausch zwischen den Wissenschaftlern ist für den Nutzer aber nur gegeben, wenn die Datenbeschaffung, angefangen von der Recherche bis hin zum Download, einfacher und weniger zeitintensiv ist bei gleicher Qualität.

Im Rahmen des Projektes EWIG wurden diverse für meteorologische Fragestellungen

Kategorien

- Data Management Plan (4)
- Forschungsdaten (13)
- Langzeitarchivierung (14)
- ohne Kategorie (5)
- Produkte und Tools (8)
- Projekte (8)
- Veranstaltungen (10)
- Veröffentlichungen (9)

Tags

Bibliographie Big Data Empfehlungen Expertengespräche Forschungsdaten Infrastruktur Hardware Hochschulen journal Kostenmodelle Lifecycle LTA DKRZ Klimadaten Marktübersicht Policy Projektergebnisse publication RDM SCAPE Strategie

kobv Kooperativer Bibliotheksverbund Berlin-Brandenburg

ZIB