Qualitätssicherung bei wenig verbreiteten Datenformaten

Martin Vogel,
Niedersächsisches Landesarchiv

Wer bin ich

- Martin Vogel
- Dipl.-Wirt.-Inf.
- Seit dem 01.05.2019 beim Niedersächsischen Landesarchiv im Projekt „Digitales Archiv“ tätig
- Zuständig für
  - SIP Bildung
  - Validierung von Daten
  - DIMAG / IngestTool

Gliederung

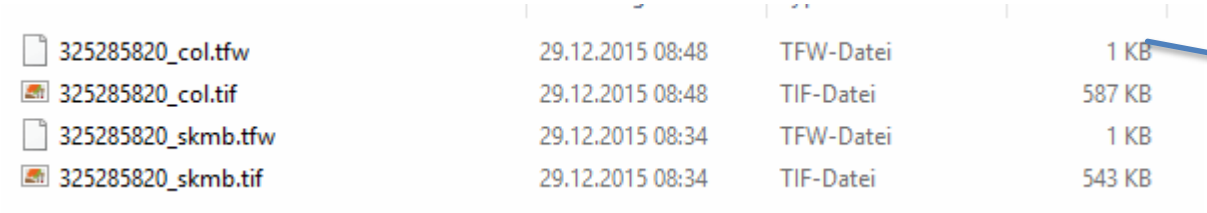
- Validierung von Daten
- Beispiel
- Eine mögliche Lösung
- Zusammenfassung / Ausblick

Validierung von Daten

- Auf Dateiebene (Syntax, Format) 
  - JHOVE
  - veraPDF
  - TreeFreeSize
  - …
- Inhaltlich (Semantik) 
  - Struktur
  - Informationsgehalt
  - Plausibilität der Daten
  - …
- Problem 
  - Es kann nur sichergestellt werden, dass Formate valide sind (z.B. PDF, TIFF), Dateien nicht leer sind usw.
  - Inhaltliche Validierung findet i.A. nicht statt
  - Nicht nutzbare Daten werden archiviert

Beispiel: Digitale Topographische Karten (DTK)

- Struktur auf Dateiebene



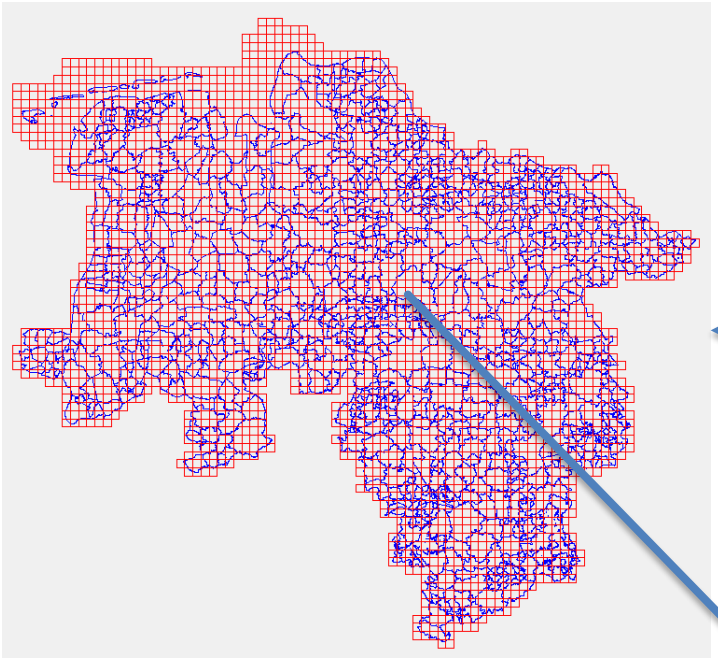
1.2500000000
0.0000000000
0.0000000000
-1.2500000000
528000.6250000000
5823999.3750000000

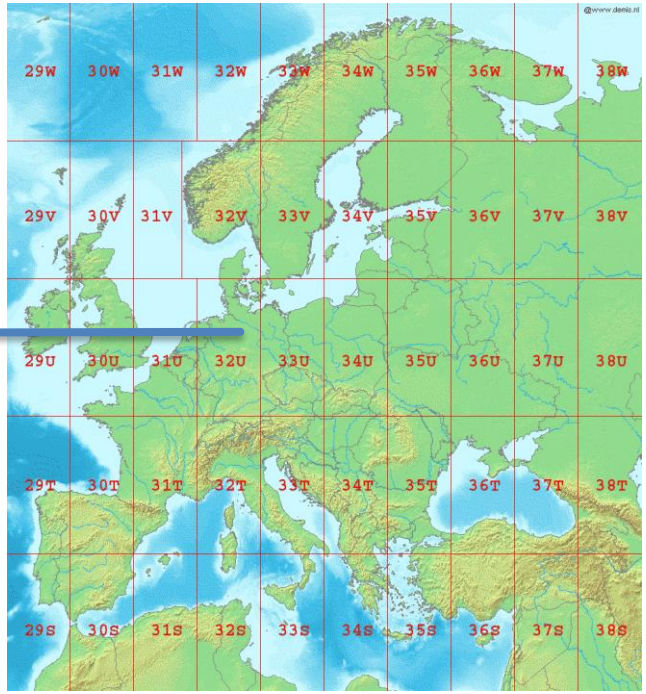
325285820_col.tfw

- Validierung Syntax:
  - Ist eine Datei leer (TreeSizeFree)
  - Sind die Tiff-Dateien valide (JHOVE)

- Überlegungen Validierung Semantik:
  - Existiert zu jeder .tfw-Datei eine .tif-Datei (Struktur)
  - Enthält jede .tfw-Datei genau 6 Zeilen (Struktur)
  - Sind die UTM-Koordinaten in den .tfw-Dateien valide (Plausibilität)

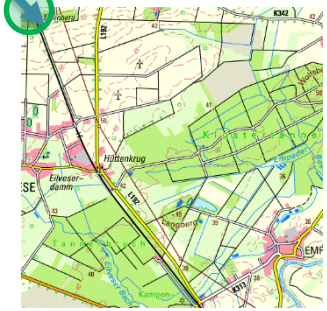
Beispiel DTK





1.2500000000
0.0000000000
0.0000000000
-1.2500000000
528000.6250000000
5823999.3750000000

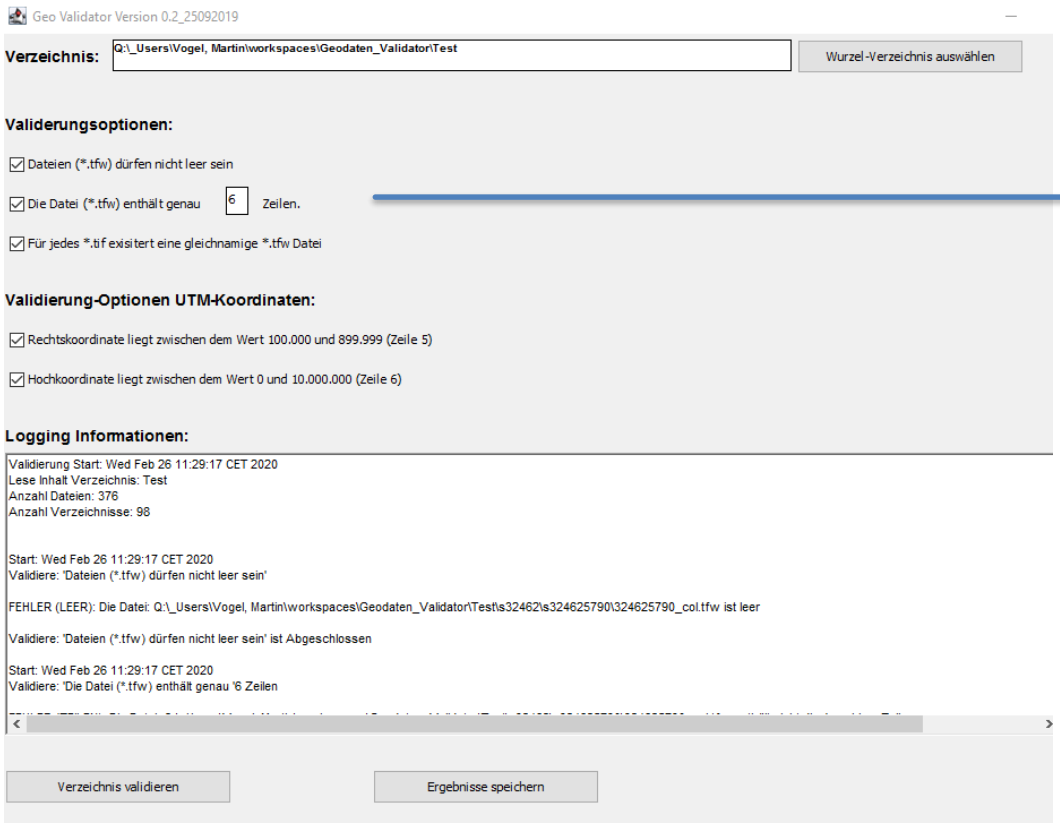
325285820_col.tfw

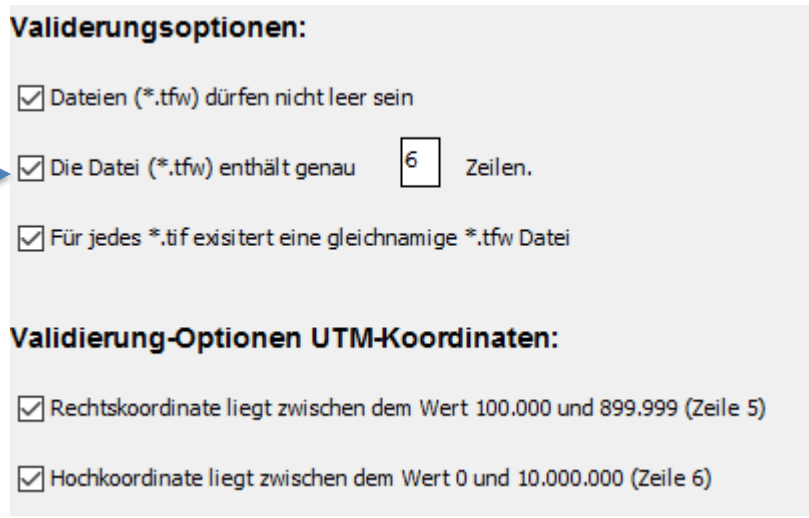


325285820_col.tif

Eine mögliche Lösung

- Eigenentwicklung mit Java



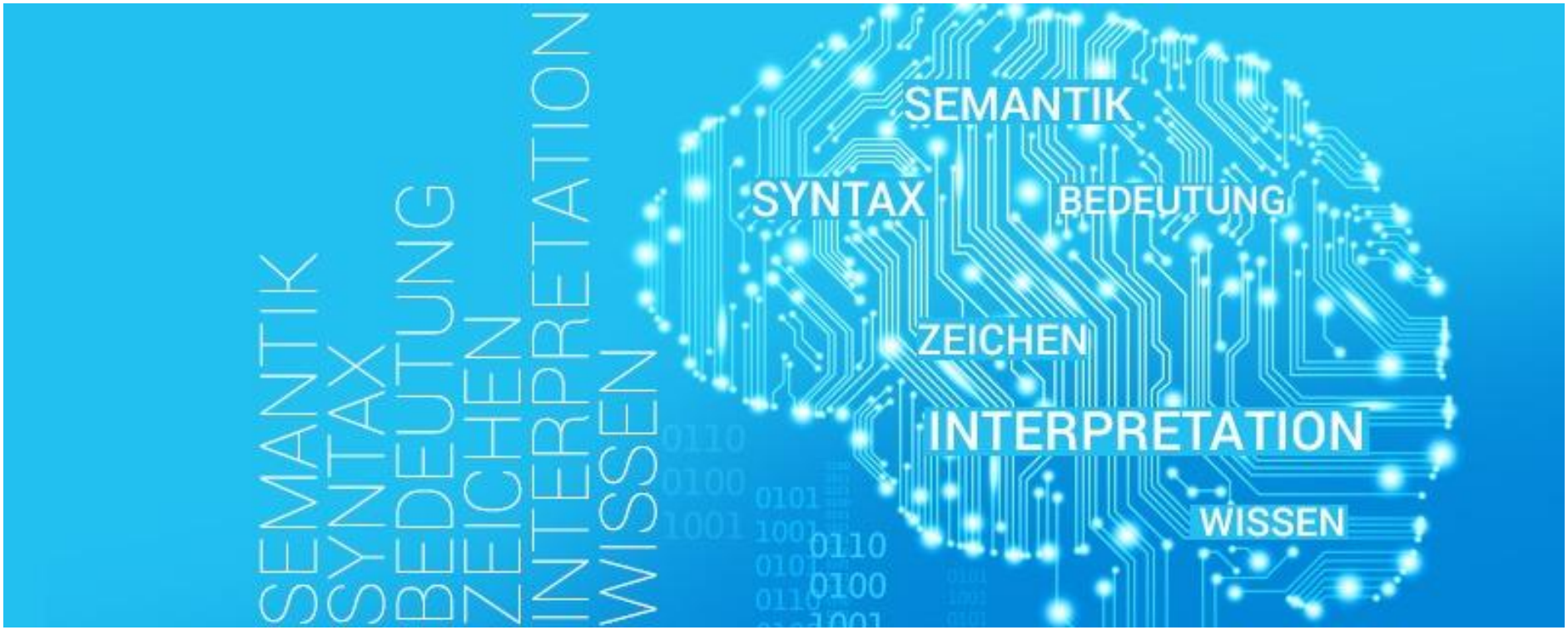


Zusammenfassung / Ausblick

- Auf Semantik kann nur validiert werden, wenn die Daten eine feste Struktur haben

- Durch die Validierung der Semantik
  - Erhöhung der Datenqualität
    - UTM Koordinaten sind valide
    - Struktur der Daten (*.tfw und *.tif existieren)
    - Struktur der *.tfw-Dateien sind valide

- Erweiterung des Validators
  - Sicherstellen, das die Koordinaten in z.B. Niedersachsen liegen
  - …


SEMANTIK
SYNTAX
BEDEUTUNG
ZEICHEN
INTERPRETATION
WISSEN
SEMANTIK
SYNTAX
BEDEUTUNG
ZEICHEN
INTERPRETATION
WISSEN