



# 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

325285820_col.tfw	29.12.2015 08:48	TFW-Datei	1 KB
325285820_col.tif	29.12.2015 08:48	TIF-Datei	587 KB
325285820_skmb.tfw	29.12.2015 08:34	TFW-Datei	1 KB
325285820_skmb.tif	29.12.2015 08:34	TIF-Datei	543 KB

```
1.2500000000  
0.0000000000  
0.0000000000  
-1.2500000000  
528000.6250000000  
5823999.3750000000
```

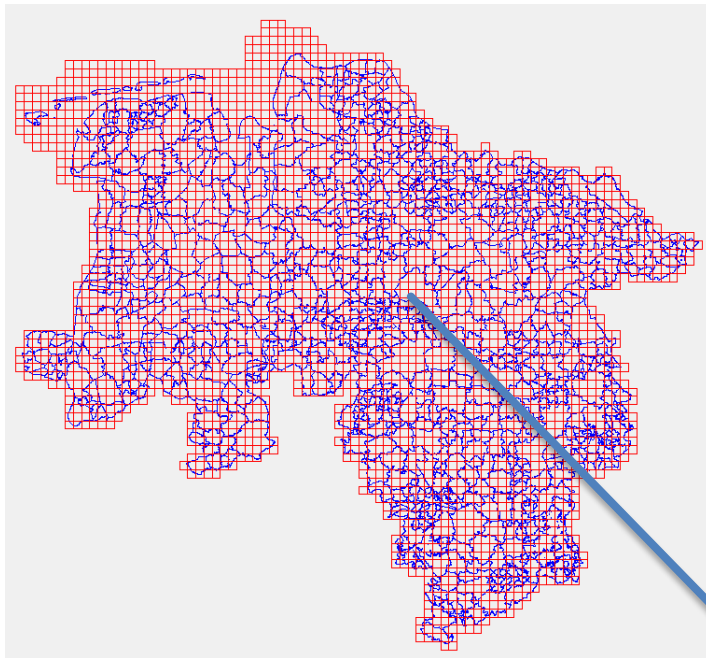
325285820\_col.tfw

- 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)

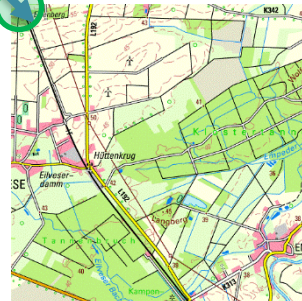


# Beispiel DTK



1.250000000  
0.000000000  
0.000000000  
-1.250000000  
**52800.625000000**  
**5823999.375000000**

325285820\_col.tfw



325285820\_col.tif



# Eine mögliche Lösung

- Eigenentwicklung mit Java

Geo Validator Version 0.2\_25092019

Verzeichnis:

**Validierungsoptionen:**

Dateien (\*.tfw) dürfen nicht leer sein

Die Datei (\*.tfw) enthält genau  Zeilen.

Für jedes \*.tif existiert eine gleichnamige \*.tfw Datei

**Validierung-Optionen UTM-Koordinaten:**

Rechtskoordinate liegt zwischen dem Wert 100.000 und 899.999 (Zeile 5)

Hochkoordinate liegt zwischen dem Wert 0 und 10.000.000 (Zeile 6)

**Logging Informationen:**

Validierung Start: Wed Feb 26 11:29:17 CET 2020  
Lese Inhalt Verzeichnis: Test  
Anzahl Dateien: 376  
Anzahl Verzeichnisse: 98

Start: Wed Feb 26 11:29:17 CET 2020  
Validiere: 'Dateien (\*.tfw) dürfen nicht leer sein'

FEHLER (LEER): Die Datei: Q:\\_Users\Vogel, Martin\workspaces\Geodaten\_Validator\Test\32462\324625790\324625790\_col.tfw ist leer

Validiere: 'Dateien (\*.tfw) dürfen nicht leer sein' ist Abgeschlossen

Start: Wed Feb 26 11:29:17 CET 2020  
Validiere: 'Die Datei (\*.tfw) enthält genau 6 Zeilen'

## Validierungsoptionen:

- Dateien (\*.tfw) dürfen nicht leer sein
- Die Datei (\*.tfw) enthält genau  Zeilen.
- Für jedes \*.tif existiert eine gleichnamige \*.tfw Datei

## Validierung-Optionen UTM-Koordinaten:

- Rechtskoordinate liegt zwischen dem Wert 100.000 und 899.999 (Zeile 5)
- Hochkoordinate liegt zwischen dem Wert 0 und 10.000.000 (Zeile 6)



# 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
  - ...





