Metadaten für die Automatisierung

Lightning Talk

Dr. Anna Kasprzik,
ZBW – Leibniz-Informationszentrum Wirtschaft
Mannheim, 03. April 2019

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Die ZBW ist Mitglied der Leibniz-Gemeinschaft.

Problemstellung

Vorgang: Automatisierung der Sacherschließung mit Machine-Learning-Methoden

Input: Diverse Metadaten in Textform bzw. Kombinationen davon
(Bsp.: Titel, Autoren-Keywords, Abstract, „Volltext“, …)

Output: Menge von Schlagworten aus dem jeweiligen kontrollierten Vokabular

Frage: An meiner Institution, was ist die genaue Menge von Ressourcen,
die für eines der in Frage kommenden Machine-Learning-Verfahren
zu einem bestimmten Zeitpunkt (Default: jetzt) zur Verfügung stehen?

Erkenntnis: Ist anhand der vorhandenen Metadaten kaum exakt ermittelbar!
Metadatenablage für die Automatisierung beginnt VOR der Automatisierung!

Welche Angaben bräuchte ich denn? – I

- Elemente: Titel, (Untertitel), freie Keywords, verbale Sacherschließung, klassifikatorische Sacherschließung, Abstract, Inhaltsverzeichnis, Textkörper, Referenzen, Glossar, Index, weitere Verzeichnisse oder Anhänge, …
- jeweils:
  - in Metadatensatz als Textstring enthalten
  - Link enthalten auf Ressource:
    - durchsuchbar
    - nicht-durchsuchbar (→ OCR + angeschlossener Workflow nötig?)
- jeweils: eigen oder fremd? (bei Links: wo gehostet?)

Welche Angaben bräuchte ich denn? – II

Lizenzrechtliche Angaben!

- Gibt es schon Abstufungen von TDM-Rechten, z.B. nach
  - Elementen, die ich verwenden darf
  - Methoden, die ich verwenden darf?
  - Art und Dauer der Speicherung eines verwendeten Korpus, die ich vornehmen darf (je nachdem, ob für Forschung oder Produktivbetrieb)?

  Wenn ja: muss alles strukturiert abgelegt werden!

- Zu jeder rechtlichen Konstellation: Anfangs- und Enddatum ihrer Gültigkeit!

Angaben nach/zu einem automatisierten Verfahren

Provenienz!

- von wem? wann? etc.
- welches Verfahren? welche Daten? welche Software? Konfidenzwert?

komplexer:

- Kombination mehrerer Verfahren
- Kombination verschiedener Automatisierungsvorgänge
  (z.B. automatisierte Beschlagwortung, dann Reasoning, Clustering …)

– ab da wird es extrem komplex… Welche Angaben sind dennoch wünschenswert? Welche sollten dem Katalogisierenden angezeigt werden, welche nicht?

(Interindexiererkonsistenz liegt generell bei 30–40%)

Fragen und Desiderate

erfordert Erweiterung des Metadatenschemas
→ verschärfte Katalogisierungsregeln – ?

Wie bilden wir das ab?

- keine optionalen Freitextfelder mehr! sondern standardisierte Codes

ABER: Wie setzen wir das um?

- Wer entwickelt und vor allem pflegt diese Codes?
  Bekommen wir eine entsprechende Lobby und Taskforce zusammen?
- In welchen Feldern soll das abgelegt werden, wie etablieren wir das?
  Auf welcher Ebene? (lokal, Verbund, national, FOLIO, Linked Open Data Cloud, ..)
- Wie vermitteln / Akzeptanz erzeugen?

[Knackpunkt – darf keine Mehrarbeit erzeugen –
Anpassung der Metadaten(schemata)
ebenfalls automatisierbar?]

Bei Interesse an Austausch bitte gerne melden bei

Anna Kasprzik

a.kasprzik@zbw.eu

040 42834-425

https://www.zbw.eu/ueber-uns/arbeitsschwerpunkte/metadatengenerierung/