Universität Augsburg
Universitätsbibliothek

Inhaltserschließung mit Culturegraph im B3Kat

Dr. Martin Völkl (AG Sacherschließung)

21.11.2024

Agenda

1. Funktionsweise von Culturegraph

„Culturegraph: Der Robocop der Inhaltserschließung“

2. Phase 1: Anreicherung von Formangaben und Schlagwortfolgen
3. Fehleranalyse und Bereinigungsaktionen
4. Phase 2: Vorbereitung der Anreicherung mit RVK-Notationen

„The Return of Culturegraph“

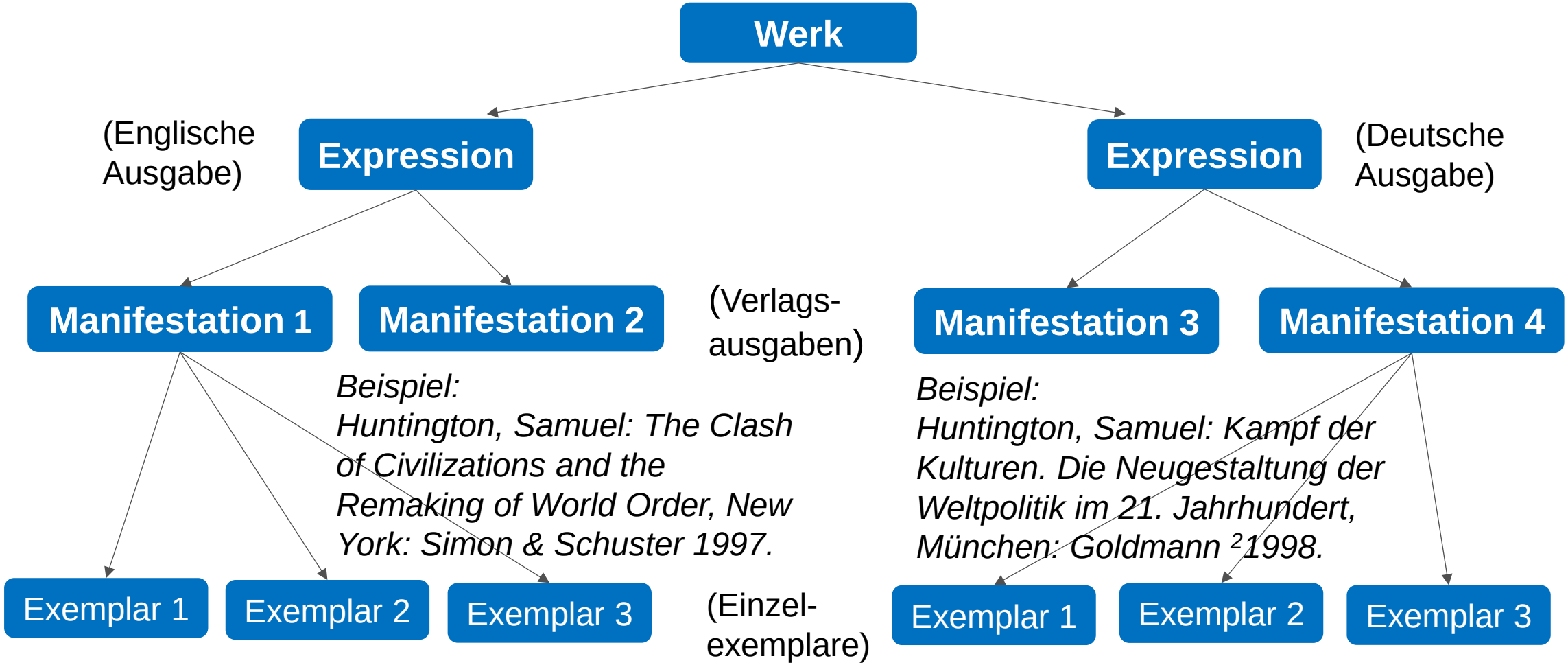
5. Zum Schluss: Wofür Culturegraph aber auch steht…

(Bild wurde erstellt mithilfe von *https://snipki.de/flux-ki-bilder-generator/*)

1. Funktionsweise von Culturegraph

- Werkclusterverfahren, mit dessen Hilfe inhaltserschließende Metadaten aller Manifestationen eines Werks miteinander abgeglichen und gegenseitig ergänzt werden können
- Culturegraph bildet auf der Grundlage der Katalogdaten Werkbündel (= „Cluster“), wobei in jedem Bündel alle Manifestationen eines Werks (⇨ Titeldatensätze im Katalog) als einzelne Bündelmitglieder enthalten sind.
- Grundprinzip von Culturegraph:
  - Inhaltserschließende Metadaten, die aufgrund der Autopsie eines vorliegenden Einzelexemplars (⇨ Buch / eBook) gewonnen werden, gelten nicht nur für die eine Manifestation (⇨ Verlagsausgabe), die durch das Exemplar repräsentiert wird, sondern auch für alle anderen Manifestationen eines Werks.
  - Abgleich der inhaltserschließenden Metadaten in allen Bündelmitgliedern
  - Ziel: Automatisierte Übertragung der inhaltserschließenden Metadaten von einer Manifestation auf alle anderen Manifestationen eines Werks (wenn dort nicht vorhanden)

1. Funktionsweise von Culturegraph



2. Phase 1: Anreicherung von Formangaben und Schlagwortfolgen

- Phase 1: Anreicherung von GND-Einzelschlagwörtern, Schlagwortfolgen und Formangaben
  - Datengrundlage: Werkbündel, die von der DNB mithilfe des Culturegraph-Verfahrens aus den Katalogdaten des B3Kat gewonnen wurden (⇨ ausschließlich die in den Titelsätzen des B3Kat-Bestandes enthaltenen Einzelschlagwörter, Schlagwortfolgen und Formangaben)
  - Umsetzung im dritten Quartal 2021
- Anreicherungsregeln:
  - Formangaben (= FA) werden nur dann im ASEQ-Feld 064a angereichert, wenn im betreffenden Titeldatensatz noch keine FA vorhanden ist.
  - Angereichert werden in den einzelnen Titelsätzen nur solche Schlagwortfolgen (= SWF) , die dort nicht bereits vorhanden sind. Die Reihenfolge der Schlagwörter in einer Schlagwortfolge ist dabei irrelevant (⇨ Permutationsprüfung).
  - Kürzere SWF, die in längeren SWF enthalten sind, werden ebenfalls nicht angereichert.

2. Phase 1: Anreicherung von Formangaben und Schlagwortfolgen

- Wenn aufgrund des Umstands, dass ein Titelsatz im B3Kat nur maximal zehn SWF aufnehmen kann, nicht alle in Frage kommenden SWF angereichert werden können, sollen diejenigen mit den meisten Schlagwörtern priorisiert werden. (Bei gleich vielen Schlagwörtern: Bevorzugung der SWF, deren Quelldatensatz mehr Besitznachweise im B3Kat aufweist.)
- Alle angereicherten FA, Einzelschlagwörter und SWF erhalten eine Provenienzkennung.

- Ergebnisse der ersten Anreicherungsphase:
  - Insgesamt wurden ca. 1,7 Millionen Print-Titeldatensätze und knapp 470.000 Titelsätze von eBooks mit inhaltserschließenden Metadaten angereichert.
  - Bei knapp 900.000 Titeldatensätzen wurden Formangaben (in ASEQ-Feld 064a) angereichert.
  - Angereichert wurden insgesamt etwa 4,1 Millionen GND-Schlagwörter oder SWF.
  - 1.004.177 angereicherte Titelsätze wiesen vorher keine inhaltliche Erschließung mit Einzelschlagwörtern oder Schlagwortfolgen auf.
  - Jede Anreicherung in einem Titelsatz wurde mit einer Provenienzangabe versehen.

2. Phase 1: Anreicherung von Formangaben und Schlagwortfolgen

Titelsätze nach der Culturegraph-Anreicherung (am Beispiel der beiden o.g. Manifestationen von Samuel Huntingtons Buch „Clash of Civilizations“ / „Kampf der Kulturen“): links die englische, rechts die deutsche Ausgabe.

Die englische Ausgabe von 1997 hatte bereits zwei Schlagwortfolgen in den Feldern 902 und 907.

Angereichert wurden hier nur zwei Schlagwortfolgen und ein Einzelschlagwort (Felder 912, 917 und 922).

Die deutsche Ausgabe hatte keine Vorerschließung, alle 5 Schlagwortfolgen wurden angereichert (daher auch mit der Provenienzangabe „cg“ für „Culturegraph“ in den Feldern mit dem Indikator „b“).

902	|s Internationaler Konflikt |9 (DE-588)416207
902	|s Zivilisation |9 (DE-588)4067906-8
902	|s Weltpolitik |9 (DE-588)4065449-7
902	|s Zukunft |9 (DE-588)4068097-6
903	|a 2134 |a 3124 |a 4123
904a	|a DE-188
907	|s Internationaler Konflikt |9 (DE-588)416207
907	|s Kulturraum |9 (DE-588)4165988-0
907	|s Internationales politisches System |9 (DE-5
907	|s Zukunft |9 (DE-588)4068097-6
908	|a 2134 |a 3124 |a 4123
909a	|a DE-188
912	|s Weltpolitik |9 (DE-588)4065449-7
912	|s Zukunft |9 (DE-588)4068097-6
912	|s Kulturkreis |9 (DE-588)4165980-6
912	|s Internationaler Konflikt |9 (DE-588)4162071-
912	|s Weltreligion |9 (DE-588)4065455-2
913	|a 21345 |a 31245 |a 41235 |a 51234
914b	|a 1@@acgwrk@@d20201028@@qDE-101@@uhttps://d-nb.info/provenance/plan#cgwrk
917	|s Weltpolitik |9 (DE-588)4065449-7
917	|s Zivilisation |9 (DE-588)4067906-8
917	|s Weltordnung |9 (DE-588)4126263-3
917	|s Kulturkonflikt |9 (DE-588)4127656-5
918	|a 2134 |a 3124 |a 4123
919b	|a 1@@acgwrk@@d20201028@@qDE-101@@uhttps://d-nb.info/provenance/plan#cgwrk
922	|s Ost-West-Konflikt |9 (DE-588)4075770-5
923	|a 2134 |a 3124 |a 4123
924b	|a 1@@acgwrk@@d20201028@@qDE-101@@uhttps://d-nb.info/provenance/plan#cgwrk

902	|s Weltpolitik |9 (DE-588)4065449-7
902	|s Zukunft |9 (DE-588)4068097-6
902	|s Kulturkreis |9 (DE-588)4165980-6
902	|s Internationaler Konflikt |9 (DE-588)41620
902	|s Weltreligion |9 (DE-588)4065455-2
903	|a 21345 |a 31245 |a 41235 |a 51234
904b	|a 1@@acgwrk@@d20201028@@qDE-101@
907	|s Internationaler Konflikt |9 (DE-588)41620
907	|s Zivilisation |9 (DE-588)4067906-8
907	|s Weltpolitik |9 (DE-588)4065449-7
907	|s Zukunft |9 (DE-588)4068097-6
908	|a 2134 |a 3124 |a 4123
909b	|a 1@@acgwrk@@d20201028@@qDE-101@@uhttps://d-nb.info/provenance/plan#cgwrk
912	|s Internationaler Konflikt |9 (DE-588)4162071-9
912	|s Kulturraum |9 (DE-588)4165988-0
912	|s Internationales politisches System |9 (DE-588)4125488-0
912	|s Zukunft |9 (DE-588)4068097-6
913	|a 2134 |a 3124 |a 4123
914b	|a 1@@acgwrk@@d20201028@@qDE-101@@uhttps://d-nb.info/provenance/plan#cgwrk
917	|s Weltpolitik |9 (DE-588)4065449-7
917	|s Zivilisation |9 (DE-588)4067906-8
917	|s Weltordnung |9 (DE-588)4126263-3
917	|s Kulturkonflikt |9 (DE-588)4127656-5
918	|a 2134 |a 3124 |a 4123
919b	|a 1@@acgwrk@@d20201028@@qDE-101@@uhttps://d-nb.info/provenance/plan#cgwrk
922	|s Ost-West-Konflikt |9 (DE-588)4075770-5
924b	|a 1@@acgwrk@@d20201028@@qDE-101@@uhttps://d-nb.info/provenance/plan#cgwrk

3. Fehleranalyse und Bereinigungsaktionen

- Die Analyse fehlerhafter Anreicherungen ergab drei verschiedene Fehlertypen:
  - Maschinell ⇨ fehlerhafte Bündelung, so dass in einem Werkcluster Manifestationen unterschiedlicher Werke enthalten waren
  - Intellektuell ⇨ nicht regelwerkskonforme Verwendung von Formangaben und Schlagwörtern, die infolge des Culturegraph-Projekts auf andere Manifestationen desselben Werks übertragen wurden
  - Strukturell ⇨ regelwerkskonforme Erschließung führt bei Bündelung der inhaltserschließenden Metadaten unterschiedlicher Manifestationen zu inhaltlich fehlerhaften Anreicherungen
- Insgesamt wurden bisher 14 Bereinigungsaktionen geplant (⇨ Rücknahme von angereicherten Formangaben und/oder Schlagwortfolgen), von denen bereits 12 durchgeführt wurden.
- Zusätzlich wurde in ASEQ-Feld 078n die Möglichkeit geschaffen, intellektuell vorgenommene Korrekturen im betreffenden Titeldatensatz zu vermerken (⇨ „cgwrk-korr“).

3. Fehleranalyse und Bereinigungsaktionen

- Beispiel für eine **Bereinigung nicht regelwerkskonformer Formangaben**
  - Bereinigung aller Titeldatensätze im B3Kat (⇨ nicht nur derjenigen, die im Culturegraph-Projekt angereichert wurden), die in ASEQ-Feld 064a sowohl die Formangabe Reisebericht, als auch - in einem weiteren ASEQ-Feld 064a - die Formangabe Tagebuch aufweisen
  - Grund für die Bereinigungsaktion: Verwendungshinweis im GND-Normdatensatz zur Formangabe Reisebericht: „Bei Reiseberichten in Tagebuchform wird nur der Formbegriff Reisebericht vergeben; ...“)
- Beispiel für die **Rücknahme der Anreicherung nicht regelwerkskonformer Schlagwortfolgen**
  - Bei historischen Quellentexten (und auch bei Comics, z.B. Asterix) war in der SWF gelegentlich die Sprache (ggf. in Verbindung mit dem Schlagwort „Übersetzung“) angegeben, in die ein historischer Quellentext (bzw. Comic) in der erschlossenen Manifestation übersetzt wurde.
  - Grund für die Bereinigungsaktion: Hier wurde nicht das Thema des Werks erschlossen, sondern die Sprache der Expression (bzw. die Übersetzung in diese Sprache)!

3. Fehleranalyse und Bereinigungsaktionen

- Beispiel für die **Rücknahme struktureller Fehler** der Culturegraph-Anreicherung **von bestimmten Formangaben**
  - Ein historischer Quellentext kann im B3Kat sowohl als Inkunabel, als auch als historisch-kritische Edition vorliegen, die gleichzeitig auch eine Hochschulschrift sein kann.
  - Problem: Alle genannten Fälle sind Manifestationen desselben Werks, sind also im selben Werkcluster enthalten.
  - Titeldatensätze dieses Bündels, die noch keine FA aufwiesen, erhielten fälschlicherweise alle drei Formangaben Inkunabel, Quelle und Hochschulschrift.
- Beispiel für die **Rücknahme struktureller Fehler** der Culturegraph-Anreicherung **von bestimmten Schlagwortfolgen**
  - Befinden sich in dem Bündel eines Werks, das eine historische Entwicklung bis zur Gegenwart schildert, verschiedene inhaltlich erweiterte Auflagen, dann bildet sich das Erscheinungsdatum der jeweiligen Auflage i.d.R. auch in der Schlagwortfolge ab.

3. Fehleranalyse und Bereinigungsaktionen

- Beispiel:

  Castles, Stephen / Miller, Mark J.: The age of migration. International population movements in the modern world, 4. Auflage, Basingstoke [u.a.] 2009.

  ⇨ Im B3Kat ist diese Manifestation nach der Culturegraph-Anreicherung mehrfach erschlossen mit derselben Schlagwortfolge $s Internationale Migration / $z Geschichte, wobei sich diese Schlagwortfolgen nur in der zeitlichen Erstreckung unterscheiden: 1945-1990, oder 1945-2001, oder 1945-2008 (also immer bis zu dem Jahr, das in der betreffenden Auflage noch behandelt wird - häufig dem Erscheinungsjahr).
- Hier wird also nicht das Werk, sondern die jeweilige Manifestation erschlossen, deren zeitliche Erstreckung sich von einer zur nächsten Auflage ändert.
- Diese inhaltliche Erschließung ist zwar regelwerkskonform, führt aber bei der Culturegraph-Anreicherung zu falschen Ergebnissen.
- Analog Fälle traten bei Katalogen, Verzeichnissen und Inventaren kunsthistorischer Sammlungen und Museen auf (⇨ erweiterte und aktualisierte Neuauflagen).

4. Phase 2: Vorbereitung der Anreicherung mit RVK-Notationen

- Phase 2 (⇨ Umsetzung geplant für 2025): Anreicherung von RVK-Notationen
- Auch hier werden die anzureichernden RVK-Notationen ausschließlich aus den Titelsätzen des B3Kat gewonnen, aus ASEQ-Feld 701g.
- Die wichtigsten, bisher formulierten Grundbedingungen für die Anreicherung:
  - Es werden nur Titeldatensätze aus Werkclustern angereichert, die noch keine RVK-Notationen enthalten (⇨ analog zu den Formangaben in Phase 1).
  - Es werden nur gültige RVK-Notationen angereichert (⇨ also nur RVK-Notationen in ASEQ-Feld 701g, nicht 701i).
  - Es wird nur bei Werkclustern angereichert, in welchen maximal 18 Manifestationen enthalten sind (⇨ hiermit wird eine Fehlerquelle aus Phase 1 minimiert).
  - Es werden maximal zehn Notationen angereichert (⇨ idealerweise nicht mehr als vier Notationen aus demselben Fachbereich, um zusätzliche Sucheinstiege aus anderen Fachbereichen der RVK zu bieten).

4. Phase 2: Vorbereitung der Anreicherung mit RVK-Notationen

- Angereichert werden soll in mehreren Teilschritten (⇨ unterschiedliche Provenienzangaben).
  - 1. Teilschritt: Angereichert wird auf der Grundlage einer Positivliste, die alle RVK-Notationen enthält, außer
    - Notationen, die in der Benennung die Begriffe „Allgemeines“ oder „Sonstige*“ enthalten
    - Notationen aus bestimmten Fachbereichen, die in der Benennung Formbegriffe enthalten, die zu fehlerhafter Anreicherung führen können (⇨ Formalstellen der RVK, z.B. mit der Benennung „Übersetzung“, „Gesamtausgabe“ oder z.T. „Quellen“ ⇨ Einzelprüfung entsprechender Systemstellen in allen Fachbereichen der RVK)
  - 2. Teilschritt: Angereichert werden nun auch RVK-Notationen, in deren Benennung die Begriffe „Allgemeines“ oder „Sonstige*“ vorkommen, wenn beim 1. Teilschritt noch keine RVK-Notation angereichert werden konnte oder wenn die angereicherten Notationen aus einem andern Fachbereich als bei Teilschritt 2 stammen.

⇨ **Vorbereitung der Anreicherung von RVK-Notationen ist aufgrund der möglichen Fehlerquellen sehr viel zeitaufwendiger und komplexer als bei Phase 1.**

5. Zum Schluss: Wofür Culturegraph aber auch steht…

Zum Wesenskern von Bibliotheken gehört ein Grundkonzept bibliothekarischer Arbeit, das sich auch im Culturegraph-Projekt manifestiert – **Kooperation:**

- Verbundübergreifend
- Beteiligung vieler Institutionen
- Beteiligung aller in der Inhaltserschließung tätigen Kolleginnen und Kollegen des B3Kat-Raums

Vielen Dank

für Ihre Aufmerksamkeit

(Bild wurde erstellt mithilfe des Bildgenerators Artguru)

Universität Augsburg
Universitätsbibliothek

Dr. Martin Völkl
Universitätsbibliothek
Universität Augsburg
martin.voelkl@bibliothek.uni-augsburg.de
www.uni-augsburg.de