Bericht Evaluierung I
AP4-2

Überprüfung des ETL-Prozesses im Projekt SoNAR (IDH)

Sina Menzel, Vivien Petras & Elena Leitner

Humboldt-Universität zu Berlin

DFKI Berlin

April 2020

Abstract

Dieser Bericht dokumentiert die Evaluierung I (AP4-2) im Projekt SoNAR (IDH). Geprüft wurde zum einen die Umsetzung des SoNAR-Datenmodells und zum anderen die Übertragung der relevanten Dateneinheiten aus dem Ausgangsdatensatz in den Zieldatensatz. Beides wurde im Ergebnis der Evaluierung als erfolgreich umgesetzt bewertet.

Inhalt

1. Ziel der Evaluierung I

Die Evaluierung I bezieht sich auf die Ergebnisse der Aufgabe 1 des Arbeitspaketes 1 (kurz: AP1-1) im Projekt SoNAR (IDH). Sie beantwortet die Frage, ob die Transformation der für SoNAR (IDH) vorliegenden Daten in ein einheitliches Datenformat sowie deren Migration in eine projektinterne Graphdatenbank über die Software neo4j[1] ohne Daten- und Potenzialverluste erreicht wurde. Als Potenzial wird die Anzahl der vorhandenen, gültigen Referenzen innerhalb der Ausgangsdatenbestände verstanden[2]. Damit werden konkret die Ergebnisse des *Extract-Transform-Load-Prozesses* (ETL) evaluiert. Ausgangsdatenbestände sind Datendumps vom Kalliope Verbund (KPE), der Zeitschriftendatenbank (ZDB), der Deutschen Nationalbibliothek (DNB) und der Gemeinsamen Normdatei (GND) aus dem Juni 2019.

Während des Prozesses des Datenmappings fand bereits ein engmaschiger, bilateraler Austausch mit AP1 statt, sodass die Herangehensweise und die daraus resultierende Architekturskizze und das Datenmodell für SoNAR (IDH) mit allen Projektpartnern abgestimmt werden konnte. Grundlage für die Transformation und Potenzialanalyse ist das projektintern beschlossene Datenmodell vom 07.11.2019[3]. Dieses Modell ist selbst nicht Bestandteil der Evaluierung I.

Der folgende Bericht stellt die Fragestellungen, Methoden und Ergebnisse der Evaluierung I dar und endet ggf. mit sich daraus ergebenden Handlungsempfehlungen für den weiteren Projektverlauf. Die Überprüfung durch AP4 fand in enger Abstimmung mit AP1 statt, sodass bereits während des Überprüfungsprozesses einige Ergänzungen an den Auszählungen vorgenommen werden konnten.

2. Quantitative Überprüfung

Dieser Abschnitt geht auf die Schritte *Extract* und *Transform* ein. Dabei wird das Vorgehen in der Datenreduktion und -transformation überprüft.

Die *quantitative Überprüfung* vergleicht daher die verarbeiteten Datenmengen und beantwortet folgende Fragen:

1. Sind alle gültigen Datensätze aus dem Ausgangsdatensatz als Knoten in die Graphdatenbank überführt?
2. Sind alle gültigen Referenzen als Kanten in die Graphdatenbank überführt?

Als “gültig” werden dabei jene Datensätze und Relationen bezeichnet, die für die automatisierte Verarbeitung geeignet sind, also mit einem permanenten und erreichbaren Link versehen (PID) und auf einen in den Ausgangsdaten ebenfalls existenten und verlinkten Zielknoten verweisend.

[1] https://neo4j.com/. Über diese Sofwarelösung sind alle Projektpartner informiert und einig, vgl. Protokoll vom 24.09.19 (Projektinterner Link: https://onedrive.live.com/view.aspx?cid=3b9129e4b3c7c3c9&page=view&resid=3B9129E4B3C7C3C9!336&parId=3B9129E4B3C7C3C9!134&authkey=!AH5z6Zyly9DgH80&app=Word)

[2] Vgl. Antragsdokument Anhang 2, S. 28: https://onedrive.live.com/?authkey=%21AH5z6Zyly9DgH80&cid=3B9129E4B3C7C3C9&id=3B9129E4B3C7C3C9%21125&parId=3B9129E4B3C7C3C9%21106&o=OneUp (Projektinterner Zugang)

[3] Vgl. https://onedrive.live.com/?authkey=%21AH5z6Zyly9DgH80&cid=3B9129E4B3C7C3C9&id=3B9129E4B3C7C3C9%21403&parId=3B9129E4B3C7C3C9%21368&o=OneUp (Projektinterner Zugang, Dokumentveröffentlichung am 13.11.19)

2.1 Vorgehen

Basis sämtlicher Evaluierungen der Teilergebnisse im Projekt SoNAR (IDH) ist die Ausgangslage der zugrunde liegenden Daten. Vorbereitend hat AP1 zu diesem Zweck Maßnahmen zur Prüfung der Konsistenz, Redundanz und Eindeutigkeit durchgeführt und dokumentiert sowie statistische Auswertungen der Daten bereitgestellt[4]. Alle Zählungen beziehen sich dabei auf das vereinbarte Datenmodell.

Für den Ausgangsdatenbestand (M) wurden darüber hinaus durch AP1 als absolute Zahlen ermittelt (vgl. Anhang B, rein informative Werte dargestellt in grau):

- M1 Anzahl der Datensätze je Datendump und insgesamt.
- M2 Anzahl der gemäß dem Datenmodell relevanten Datenfelder je Datendump. M2 beinhaltet dabei keine Datensatz-Dubletten, aber zählt ungültige Datenfelder und Datenfeld-Dubletten, wie in M5 und M7 ausdifferenziert. Außerdem fasst M2 Datenfelder mit ein, die implizite Entitäten aus Zeitausdrücken enthalten (Chron-Terme). Diese ergeben sich aus dem Feld 548 im GND-Datensatz.
- M3 Anzahl der gültigen Referenzen je Datenfeld und Datendump ohne Dubletten auf Ebene der Datensätze oder Datenfelder. M3 beinhaltet damit allein diejenigen Referenzen, die in die Zieldatenbank übertragen werden sollen (Token).
- M4 Anzahl der gemäß M3 referenzierten, gültigen Entitäten je Entitätentyp und Datendump (Types).
- M5 Anzahl der ungültigen Referenzen je Datendump auf Ebene eines Datenfeldes nach Fehlertyp.
- M6 Anzahl der Dubletten je Datendump auf Ebene eines Datensatzes. Konkret zählt M6 die zur DNB übertragenen ZDB-Datensätze im SoNAR-Datenpool.
- M7 Anzahl der Dubletten je Datendump auf Ebene eines Datenfeldes/Unterfeldes.

Für Daten des Zielsystems (N), der Graphdatenbank neo4j, wurde durch AP1 in absoluten Zahlen ermittelt:

- N1 Anzahl der Knoten insgesamt.
- N2 Anzahl der Knoten je Datendump.
- N3 Anzahl der Knoten je Entitätentyp.
- N4 Anzahl der Knoten je Relationentyp und Datendump.
- N5 Anzahl der Kanten insgesamt.
- N6 Anzahl der Kanten je Datendump.
- N7 Anzahl der Kanten je Relationentyp.
- N8 Anzahl der Kanten je Relationentyp und Datendump.

Die Werte für M4 sowie N3, N4, N7 und N8 (dargestellt in grau) dienen der Information und Dokumentation bzw. der Ausdifferenzierung der Gesamtzahlen. Alle anderen Werte sind maßgeblich für die folgende Überprüfung.

[4] Vgl. https://onedrive.live.com/view.aspx?cid=3b9129e4b3c7c3c9&page=view&resid=3B9129E4B3C7C3C9!336&parId=3B9129E4B3C7C3C9!134&authkey=!AH5z6Zyly9DgH80&app=Word (Projektinterner Zugang)

2.2 Ergebnisse[5]

Knoten gesamt

Die im Projekt SoNAR zur Verfügung stehende Datenmenge aus den Datendumps stammen aus Quellen, die z.T. untereinander einen bilateralen Datenaustausch pflegen. So konnte im Vorfeld festgestellt werden, dass ein signifikanter Anteil der Datensätze aus der Deutschen Nationalbibliothek (DNB) mit Datensätzen aus der Zeitschriftendatenbank (ZDB) übereinstimmt. Diese Dubletten wurden ausgelesen und nicht in die Zieldatenbank übertragen. Der Umfang ist in M6 festgehalten.

Um festzustellen, ob alle gültigen Datensätze aus den einzelnen Datendumps als Knoten in die Graphdatenbank überführt wurden, wurde zunächst die Differenz der Gesamtzahl der Dubletten (M6) und der gesamten Datensätze aus den Ausgangsdaten (M1gesamt) gebildet.

Diese Zahl wurde anschließend durch jene Knoten und Kanten ergänzt, die nicht aus Datensätzen in die Zieldatenbank, sondern implizit aus Referenzen generiert werden. Hierbei handelt es sich zum einen um ISIL-Terme[6], die innerhalb von Datensätzen verzeichnet sind und als Kanten auf Körperschaften referieren (N2Isil), zum anderen um Zeitterme, die dem Feld 548 in GND-Datensätzen entnommen sind (N2Chron).[7] Diese Zeitterme werden als Knoten in der Zieldatenbank eingepflegt. Grundlage ist die Annahme, dass sowohl ISIL-Terme und Zeitausdrücke ausgehende Referenzen darstellen.

Rechnung: **(M1gesamt - M6) + N2Isil + N2Chron = N1**

Zählung	**Erläuterung**	**Operation**	**Wert**
M1	Anzahl der Datensätze insgesamt (gültig+ungültig)	Gegeben	34.516.127
M6	Anzahl der Dubletten aus DNB-Datendump	Subtraktion	541.840
	Zwischenergebnis I: M1ges-M6	Berechnet	33.974.287
N2Isil	Kanten implizit = Anzahl *RelationToIsilTerm*	Gegeben	611
N2Chron	Knoten implizit = Anzahl *ChronTerm*	Gegeben	537.054
	Zwischenergebnis II: N2Isil+N2Chron	Addition	537.665
	Zwischenergebnis III: Addition der Zwischenergebnisse	Addition	34.511.952
N1	Anzahl der Knoten in neo4j insgesamt	Gegeben	34.511.952
	Übereinstimmung Zwischenergebnis III und N1?	Vergleich	Ja

Fazit: Alle gültigen Datensätze aus dem Ausgangsdatensatz sind als Knoten in die Graphdatenbank überführt. Eingefasst sind hier sowohl explizite Knoten, als auch implizite Knoten.

[5] Die Werte der folgenden Rechnungen beziehen sich auf die Angaben in Anhang A.
[6] Internationale Standardkennzeichen für Bibliotheken und verwandte Einrichtungen, vgl. https://sigel.staatsbibliothek-berlin.de/vergabe/isil/
[7] Die Einbeziehung der ISIL DE-588-Terme und der Zeitterme als implizite Knoten wurde am 07.11.20 beschlossen, vgl. Kapitel 2.1 im Datenmodell.

Kanten gesamt und nach Datendump

Relationen zwischen den Entitäten werden für den SoNAR-Datenpool gemäß dem Datenmodell aus vordefinierten Datenfeldern gezogen. Auch auf Ebene des Datenfeldes wurden dabei im Vorfeld Fehlerquellen festgestellt, die eine erfolgreiche Übertragung bestimmter Datenfelder als Kanten in die Zieldatenbank verhindern, diese sind in M5 dokumentiert[8]. Darüber hinaus gab es redundante Informationen über Relationen innerhalb einzelner Datensätze, die in M7 festgehalten wurden.

M3 schließt diese fehlerhaften und redundanten Datenfelder aus und führt somit alle Relationen auf, die in die Zieldatenbank übertragen werden sollen. Um festzustellen, ob alle gültigen Relationen aus den einzelnen Datendumps als Kanten in die Zieldatenbank überführt wurden, konnten daher die Werte für M3 mit den Werten für N5 abgeglichen werden (Rechnung I). Zur Gegenprüfung wurde zusätzlich die Rechnung über den Wert M2 durchgeführt, der die ungültigen Referenzen und die genaue Aufteilung nach Datendump aufzeigt (Rechnung II). Bei Rechnung II werden dabei die impliziten Knoten zu Körperschaften (ISIL-Terme) addiert, die impliziten Knoten zu Zeitausdrücken sind bereits in M2 mit eingefasst (Chron-Terme).

Rechnung I: **M3ges = N5**

Zählung	Erläuterung	Operation	Wert
M3	Anzahl der gültigen Datensätze insgesamt	Gegeben	98.530.160
N5	Anzahl der Kanten insgesamt	Gegeben	98.530.160
	Übereinstimmung M3ges und N5?	Vergleich	Ja

Rechnung II: **(M2 – M5 – M7) + M3Isil = N6**

Zählung	Erläuterung	Operation	GND	DNB ohne ZDB	ZDB	KPE	Gesamt
M2	Anzahl der relevanten Datenfelder (gültig+ungültig)	Gegeben	18.118.006	31.536.618	2.858.921	12.543.306	65.056.851
M5	Anzahl der fehlerhaften Datenfelder (ungültig)	Substraktion	1.929.672	19.807.848	58.356	192.757	21.988.633
M7	Anzahl der Dubletten aus Datenfeldern (Dublette)	Substraktion	1.800	92.290	881	0	94.971
	Zwischenergebnis I: M2-(M5+M7) = Anzahl der gültigen Datenfelder (explizit)	Berechnet	16.186.534	11.636.480	2.799.684	12.350.549	42.973.247
M3	Kanten implizit = Anzahl *RelationToIsilTerm* nach Datendump	Addition	16.590.094	31.782.859	2.856.175	4.327.785	55.556.913
	Zwischenergebnis II: Anzahl der gültigen Datenfelder (explizit + implizit)	Berechnet	32.776.628	43.419.339	5.655.859	16.678.334	98.530.160
N6	Kanten in neo4j nach Datendump (explizit + implizit)	Gegeben	32.776.628	43.419.339	5.655.859	16.678.334	98.530.160
	Übereinstimmung Zwischenergebnis II und N6?	Vergleich	Ja	Ja	Ja	Ja	Ja

[8] Für eine detaillierte Fehlerbeschreibung vgl. auch die Dokumentation vom 26.09.2019: https://onedrive.live.com/edit.aspx?cid=3b9129e4b3c7c3c9&page=view&resid=3B9129E4B3C7C3C9!350&parId=3B9129E4B3C7C3C9!226&authkey=!AH5z6Zyly9DgH80&app=Word (Projektinterner Zugang)

Fazit: Alle gültigen Relationen aus dem Ausgangsdatensatz sind als Kanten in die Graphdatenbank überführt. Eingefasst sind hier sowohl explizite Kanten, als auch implizite - also abgeleitete - Kanten (RelationToIsilTerm).

3. Qualitative Überprüfung

In diesem Kapitel wird die Umsetzung des Datenmodells im Ladeprozess qualitativ überprüft. Betrachtet wird daher der *Load*-Aspekt mit Blick auf die Zieldatenbank. Für die vordefinierten Eigenschaften muss in den normalisierten Daten weiterhin eine Entsprechung vorliegen.

3.1 Vorgehen

Im Datenmodell werden drei verschiedene Instanzen definiert: Erstens die möglichen Entitätentypen, zweitens die Relationen, die zwischen diesen Entitätentypen möglich sind und drittens die Merkmale, die den Entitäten und Relationen eigen sein können. Die *qualitative Überprüfung* vergleicht die im Datenmodell spezifizierten Instanzen mit denen in der Zieldatenbank integrierten. Dabei wird folgende Frage beantwortet:

3. Sind alle im Datenmodell spezifizierten Entitätentypen, Relationentypen und die jeweiligen Merkmale von der Transformation - und nur diese - in der Zieldatenbank neo4j berücksichtigt und entsprechend bezeichnet?

Es handelt sich dabei um eine manuelle Überprüfung der Metadaten aus der Zieldatenbank. Die durch AP1 bereitgestellten transformierten Daten (Format: graphml) werden dabei lokal in neo4j Desktop 1.2.2 übertragen (v. 3.5.8). Durch Abfragen in der Datenbank wird anschließend ein Abgleich mit dem Datenmodell möglich.[9]

3.2 Ergebnisse[10]

Knotentypen

Laut Datenmodell sind neun Entitätentypen vorgesehen. Voraussetzung ist dabei ein vorhandener eindeutiger Identifikator `"EntityId"` im Normdatensatz der GND.

Abfrage in neo4j: `call db.labels();`

Ausgabe: `"CorpName", "GeoName", "MeetName", "PerName", "TopicTerm", "UniTitle", "ChronTerm", "IsilTerm", "Resource"`

Fazit: Sämtliche Entsprechungen der im Datenmodell definierten Entitätentypen - und nur diese - sind in den transformierten Datensatz übertragen.

[9] Dieser Prozess ist für alle Projektpartner reproduzierbar. Eine entsprechende Installataionsanleitung für neo4j wurde am 24.01.2020 durch Elena Leitner über die projektinterne Mailingliste bereitgestellt. Das Datenset gibt es unter AlleDaten_v.2: https://onedrive.live.com/?authkey=%21AH5z6Zyly9DgH80&id=3B9129E4B3C7C3C9%21431&cid=3B9129E4B3C7C3C9 (Projektinterner Link)

[10] Die hier beschriebenen Inhalte des Datenmodells sind in Anhang A, S.1 einzusehen.

Kantentypen

Laut Datenmodell sind neun Relationentypen definiert, die jeweils auf einen Entitätentyp verweisen (vgl. Anhang A, S.). Voraussetzung ist dabei ein vorhandener eindeutiger Identifikator `"ResourceId"` im Metadatensatz.

Abfrage in neo4j: `match (n)-[r]-() return distinct type(r)`

Ausgabe: `"RelationToIsil", "RelationToTopicTerm", "RelationToMeetName", "RelationToPerName", "RelationToChronTerm", "RelationToUniTitle", "RelationToGeoName", "RelationToResource", "RelationToCorpName"`

Fazit: Sämtliche Entsprechungen der im Datenmodell definierten Relationentypen - und nur diese - sind in den transformierten Datensatz übertragen.

Merkmale (Property Keys)

Den neun Entitätentypen sind laut Datenmodell sieben allgemeine Merkmale zugeordnet. Neben diesen allgemeinen Merkmalen sind allen Entitätentypen im Datenmodell eine oder mehrere spezifische Merkmale zugeordnet. Insgesamt handelt es sich um 37 allgemeine und spezifische Merkmale für Entitäten.

Darüber hinaus sind auch den Relationentypen fünf mögliche Merkmale zugeordnet.

Abfrage I in neo4j (alle Merkmale): `CALL db.propertyKeys`

Ausgabe:

```
"id", "CorpDateApprox", "CorpInfo", "CorpSubUnit", "EntityGenType",
"EntityId", "EntityName", "EntityOldId", "EntitySpecType", "EntityUri",
"EntityVariantName", "GeoArea", "GeoGenSubdiv", "GeoInfo", "MeetDate",
"MeetInfo", "MeetPlace", "MeetSubUnit", "PerDateApprox", "PerDateStrict",
"PerGender", "TopicGenSubdiv", "TopicInfo", "WorkCreator", "WorkDate",
"WorkLang", "WorkMedium", "RelationTypeAddInfo", "RelationSource",
"RelationSourceType", "RelationTempValidity", "ResourceCreator",
"ResourceGenre", "ResourceId", "ResourceLang", "ResourcePublDate",
"ResourcePublPlace", "ResourceTitle", "ResourceUri"
```

Abfrage II in neo4j (Merkmale der Entitätentypen):

```
MATCH(n)
WITH LABELS(n) AS labels , KEYS(n) AS keys
UNWIND labels AS label
UNWIND keys AS key
RETURN DISTINCT label, COLLECT(DISTINCT key) AS props
ORDER BY label
```

Ausgabe:

label	props
"ChronTerm"	["id" ,"EntityName"]
"CorpName"	["EntityVariantName", "EntityUri", "EntityName", "EntityId", "CorpSubUnit", "EntityGenType", "EntityOldId", "EntitySpecType", "id", "CorpDateApprox", "CorpInfo"]
"GeoName"	["GeoInfo", "EntityVariantName", "GeoArea", "GeoGenSubdiv", "EntityName", "EntitySpecType", "EntityId", "EntityUri", "id", "EntityGenType", "EntityOldId"]
"IsilTerm"	["id", "EntityName"]
"MeetName"	["MeetDate", "MeetInfo", "MeetPlace", "MeetSubUnit", "EntityUri", "EntityVariantName", "EntityId", "EntityOldId", "EntitySpecType", "id", "EntityGenType", "EntityName"]
"PerName"	["PerGender", "EntityVariantName", "PerDateApprox", "PerDateStrict", "EntityUri", "EntityName", "EntityId", "EntitySpecType", "id", "EntityGenType", "EntityOldId"]
"Resource"	["ResourceCreator", "ResourceTitle", "ResourcePublDate", "ResourcePublPlace", "ResourceUri", "ResourceId", "ResourceLang", "id", "ResourceGenre"]
"TopicTerm"	["TopicInfo", "EntityUri", "EntityName", "EntityVariantName", "TopicGenSubdiv", "EntityId", "EntitySpecType", "id", "EntityGenType", "EntityOldId"]
"UniTitle"	["WorkMedium", "WorkCreator", "WorkDate", "WorkLang", "EntityUri", "EntityName", "EntityVariantName", "EntityId", "EntitySpecType", "id", "EntityGenType", "EntityOldId"]

Abfrage III in neo4j (Merkmale der Relationentypen):

```
MATCH ()-[m]->()
WITH TYPE(m) AS type, KEYS(m) AS keys
UNWIND keys AS key
RETURN DISTINCT type, COLLECT(DISTINCT key) AS props
ORDER BY type
```

Ausgabe:

type	props
"RelationToChronTerm"	["RelationSourceType", "RelationTypeAddInfo", "RelationSource", "RelationTempValidity"]
"RelationToCorpName"	["RelationTypeAddInfo", "RelationSource", "RelationSourceType", "RelationTempValidity"]
"RelationToGeoName"	["RelationSourceType", "RelationTypeAddInfo", "RelationSource", "RelationTempValidity"]
"RelationToIsil"	["RelationTypeAddInfo"]
"RelationToMeetName"	["RelationTypeAddInfo", "RelationSource", "RelationSourceType", "RelationTempValidity"]
"RelationToPerName"	["RelationTypeAddInfo", "RelationSource", "RelationSourceType", "RelationTempValidity"]
"RelationToResource"	["RelationSource", "RelationSourceType", "RelationTempValidity", "RelationTypeAddInfo"]
"RelationToTopicTerm"	["RelationSourceType", "RelationTypeAddInfo", "RelationSource", "RelationTempValidity"]
"RelationToUniTitle"	["RelationTypeAddInfo", "RelationSource", "RelationSourceType", "RelationTempValidity"]

Fazit: Sämtliche Entsprechungen der im Datenmodell definierten allgemeinen und spezifischen Merkmale der Entitäten - und nur diese - sind in den transformierten Datensatz übertragen (Abfrage I). Diese Merkmale sind dem Datenmodell entsprechend den Entitäten und Relationen zugeordnet (Abfrage II und III). Das Merkmal `"id"` ist dabei nicht im Datenmodell definiert, sondern ein obligatorisches Merkmal im Format graphml zum Zwecke der Unterscheidung von Entiäten (node ID) und Relationen (edge ID).

4. Zusammenfassung der Ergebnisse

Frage 1: Sind alle gültigen Datensätze aus dem Ausgangsdatensatz als Knoten in die Graphdatenbank überführt?

Ergebnis: Ja. Die Zahl der gültigen Datensätze, sowie implizit abgeleitete Knoten (ISIL DE-588-Terme und Zeitterme) stimmt mit der Zahl der Knoten in der Zieldatenbank neo4j überein. Die Projektpartner in SoNAR-Projekt können somit in der Erprobung auf 34.511.952 verarbeitete Entitäten zurückgreifen.

Frage 2: Sind alle gültigen Referenzen als Kanten in die Graphdatenbank überführt?

Ergebnis: Ja. Die Zahl der gültigen Relationen in den Ausgangsdaten stimmt mit der Zahl der Kanten in der Zieldatenbank neo4j überein. Die Projektpartner in SoNAR-Projekt können somit in der Erprobung auf 98.530.160 Verknüpfungen zwischen den Entitäten zurückgreifen.

Frage 3: Sind alle im Datenmodell spezifizierten Entitätentypen, Relationentypen und die jeweiligen Merkmale von der Transformation - und nur diese - in der Graphdatenbank neo4j berücksichtigt und entsprechend bezeichnet?

Ergebnis: Ja. Die qualitative Überprüfung der Umsetzung des Datenmodells ergab keine Unstimmigkeiten.

Alle Überprüfungen sind jederzeit durch alle Projektangehörigen reproduzierbar.

5. Fazit und Handlungsbedarf

Die Extraktion, Transformation und das Laden der Ausgangsdaten wurde gemäß des Datenmodells ausgeführt. Projektintern ist daher kein Handlungsbedarf festzustellen, die Erprobung kann somit mit dem Datensatz zur Übertragung in neo4j vom 17.03.2020 fortgesetzt werden[11]. Mögliche Änderungen im SoNAR-Datenpool oder im Datenmodell würden eine erneute Überprüfung der geänderten Komponenten erfordern.

Ferner konnte über den Projektrahmen hinaus Handlungsbedarf bezüglich der Metadatenqualität festgestellt werden. Dabei handelt es sich vor allem um ungültige Datensätze und Referenzen, also Knoten und Kanten, die aufgrund fehlerhafter Metadaten nicht verarbeitet werden konnten. Umfang und Fehlertyp sind durch AP1 in M5-M7 dokumentiert und sollten den bereitstellenden Institutionen der Datendumps rückgemeldet werden.

[11] Projektinterner Zugang: https://onedrive.live.com/?authkey=%21AH5z6Zyly9DgH80&id=3B9129E4B3C7C3C9%21431&cid=3B9129E4B3C7C3C

Anhang A: Datenmodell

(Verf. Elena Leitner, Stand 24.04.20)

Entitäten

1. `PerName` von der GND
2. `CorpName` von der GND
3. `MeetName` von der GND
4. `UniTitle` von der GND
5. `TopicTerm` von der GND
6. `GeoName` von der GND
7. `ChronTerm` von der GND (aus Relationen)
8. `IsilTerm` von einer Liste mit ISILs
9. `Resource` von der DNB, ZDB, KPE

Entitäten GND

Allgemeine Merkmale

1. Identifikator (ISIL DE-588) `EntityId`
2. gelöschte Identifikatoren `EntityOldId`
3. URI `EntityUri`
4. Entitätentyp (gndgen) `EntityGenType`
5. Entitätentyp (gndspec) `EntitySpecType`
6. Entitätenname `EntityName`
7. Andere Entitätennamen `EntityVariantName`

Spezifische Merkmale

1. **Personen** `PerName`
    - Geschlecht `PerGender`
    - Lebensdaten `PerDateApprox`
    - exakte Lebensdaten `PerDateStrict`
2. **Körperschaften** `CorpName`
    - Entstehungsdaten `CorpDateApprox`
    - Untergeordnete Körperschaft `CorpSubUnit`
    - Sonstige Informationen `CorpInfo`
3. **Kongresse** `MeetName`
    - Ort des Kongresses `MeetPlace`
    - Datum des Kongresses `MeetDate`
    - Untergeordnete Einheit `MeetSubUnit`
    - Sonstige Informationen `MeetInfo`

4. **Werke** `UniTitle`
    - Beachte: Titel `EntityName`
    - Verfasser/Urheber `WorkCreator`
    - Erscheinungsjahr `WorkDate`
    - Medium `WorkMedium`
    - Sprache eines Werkes `WorkLang`
5. **Sachbegriffe** `TopicTerm`
    - Sonstige Informationen `TopicInfo`
    - Allgemeine Unterteilung `TopicGenSubdiv`
6. **Geografika** `GeoName`
    - Sonstige Informationen `GeoInfo`
    - Allgemeine Unterteilung `GeoGenSubdiv`
    - Code für geografische Gebiete `GeoArea`
7. **Zeitausdrücke** `ChronTerm`
    - Zeitausdruck `EntityName`
8. **ISIL-Einrichtungen** `IsilTerm`
    - Einrichtung `EntityName`

Ressourcen DNB, KPE

Merkmale der Ressourcen

1. Identifikator (ISIL DE-611, DE-101 ...) `ResourceId`
2. ~~alte/ungültige Identifikatoren ResourceOldId~~ Es gibt keine veraltenen Identifikatoren in Titeldaten (für weitere Informationen s. MARC21-Dokumentation bzw. Statistiken)
3. Titel `ResourceTitle`
4. Bestandsbildner (Person, Körperschaft) `ResourceCreator`
   Erscheinungsjahr (-verlauf) `ResourcePublDate`
5. Erscheinungsort `ResourcePublPlace`
6. Sprache `ResourceLang`
7. Gattung `ResourceGenre`
8. Link zur Ressource `ResourceUri`

Relationen

Attribute der Relationen:

1. Identifikator `EntityId` des Zielknoten (implizit)

2. Relationentyp `RelationType`, der den prototypischen Entitätentypen entspricht:

- `RelationToPerName`
- `RelationToCorpName`
- `RelationToMeetName`
- `RelationToUniTitle`
- `RelationToTopicTerm`
- `RelationToGeoName`
- `RelationToChronTerm`
- `RelationToIsilTerm`
- `RelationToResource`

3. Quelle einer Relation `RelationSource`

- GND
- DNB (entschieden, ZDB-Dubletten auszufiltern)
- KPE

4. Relationentyp `RelationSourceType`, der in Quelldaten entsprechend klassifiziert wird. Dazu zählt man folgende Typen:

- im GND-Datendump sind es Typen aus der GND-Ontologie
- im DNB- und ZDB-Datendump sind es mit Codes verschlüsselte Typen, genannt Relators, die für Relationen zu der GND-Entitäten typisch sind. Für Relationen zu Ressourcen aus dem DNB- und ZDB-Datendump sind je nach Datenfeld `SupplementToMainTitle`, `MainTitleToSupplement`, `LangVariant`, `ManifestLevel`, `Predecessor`, `Successor` vordefiniert.
- im KPE-Datendump sind es LevelToLevel-Typen, also `CollectionToClass`, `ClassToClass`, `ClassToItem` etc., die automatisch generiert werden.

5. Zusätzliche Information zu einer Relation `RelationTypeAddInfo`:

- im GND-Datendump sind es zusätzliche Information bezüglich Relationen aus dem Unterfeld `9v:`
- im DNB- und ZDB-Datendump sind es Beziehungskennzeichen aus dem Unterfeld `a` oder Beziehungsart bezüglich eines fortlaufenden Sammelwerks aus dem Unterfeld `i`
- im KPE-Datendump ist dieses Attribut nicht besetzt

6. Zeitliche Gültigkeit einer Relationen `RelationTypeValidity`

- im GND-Datendump sind es zeitliche Gültigkeit der Relationen aus dem Unterfeld `9Z:`
- im DNB- und ZDB-Datendump sind es sind es zeitliche Gültigkeit der Relationen je nach Datenfeld aus den Unterfeldern `b` oder `n`
- im KPE-Datendump ist dieses Attribut nicht besetzt.

Kongruenz bezüglich Feldern und Relationen

MARC21 GND

- 500 = `RelationToPerName`
- 510 = `RelationToCorpName`
- 511 = `RelationToMeetName`
- 530 = `RelationToUniTitle`
- 548 = `RelationToChronTerm`
- 550 = `RelationToTopicTerm`
- 551 = `RelationToGeoName`
- Präfix von 035 = `RelationToIsilTerm`

MARC21 DNB, ZDB

- 100, 700 = `RelationToPerName`
- 110, 710 = `RelationToCorpName`
- 111, 711 = `RelationToMeetName`
- 130, 730 = `RelationToUniTitle`
- 751 = `RelationToGeoName`
- 770 = `RelationToResource` (`RelationType` = SupplementToMainTitle)
- 772 = `RelationToResource` (`RelationType` = MainTitleToSupplement)
- 775 = `RelationToResource` (`RelationType` = LangVariant)
- 776 = `RelationToResource` (`RelationType` = ManifestLevel)
- 780 = `RelationToResource` (`RelationType` = Predecessor)
- 785 = `RelationToResource` (`RelationType` = Successor)
- Präfix von 035 = `RelationToIsilTerm`

EAD KPE

- controlaccess/persname = `RelationToPerName`
- controlaccess/corpname = `RelationToCorpName`
- controlaccess/subject = `RelationToTopicTerm`
- controlaccess/geogname = `RelationToGeoName`
- archdesc@level|c@level = `RelationToResource`
- archdesc/did/repository/corpname@authfilenumber (@source = 'ISIL') = `RelationToIsilTerm`

Anhang B: Zählungen

(Verf. Elena Leitner, Stand: 28.04.20)

M1 Anzahl der Datensätze* je Datendump

Datendump	Anzahl
GND	8.295.047
DNB	19.926.573
ZDB	1.908.334
KPE	4.386.173
ZDB-Dubletten	541.840
M1ges	35.057.967

*Ein Record entspricht einem Datensatz. IsilTerm und ChronTerm sind nicht in Records, sondern in Datenfeldern kodiert. So werden diese Entitätentypen hier nicht beschrieben. IsilTerm und ChronTerm sind in M4 gezählt.

M2 Anzahl der relevanten Datenfelder (nach Datenmodell) je Datendump

Datendumps	**Felder**										
	1XX	*500*	*510*	*511*	*530*	*550*	*551*	*548*			alle Referenzen zu Normdaten
	EntityName	*RelationTo PerName*	*RelationTo CorpName*	*RelationTo MeetName*	*RelationTo UniTitle*	*RelationTo TopicTerm*	*RelationTo GeoName*	*RelationTo ChronTerm*			
Personen	5.087.660	418.986	728.633	226	2.303	5.346.145	2.149.797	4.280.577			12.926.667
Körperschaften	1.487.711	14.863	676.128	433	2.323	305.986	1.122.308	257.432			2.379.473
Kongresse	814.044	917	37.862	40.747	12	32.562	712.906	672.480			1.497.486
Werke	385.300	387.861	26.603	311	35.327	63.977	25.007	182.122			721.208
Sachbegriffe	212.135	2.906	3.670	18	133	211.654	12.176	7.315			237.872
Geografika	308.197	9.184	3.423	24	27	115.747	179.360	47.535			355.300
GND	8.295.047	834.717	1.476.319	41.759	40.125	6.076.071	4.201.554	5.447.461			18.118.006
		100	*700*	*110*	*710*	*111*	*711*	*130*	*730*	*751*	alle Referenzen zu Normdaten
		RelationTo PerName	*RelationTo PerName*	*RelationTo CorpName*	*RelationTo CorpName*	*RelationTo MeetName*	*RelationTo MeetName*	*RelationTo UniTitle*	*RelationTo UniTitle*	*RelationTo GeoName*	
ZDB	1.908.334	837	21.639	636.950	421.080	38.779	6.390	3.527	9	201.252	1.330.463
	245	*770*	*772*	*775*	*776*	*780*	*785*				alle Referenzen zu Titeldaten
	ResourceName	*RelationTo Resource*	*RelationTo Resource*	*RelationTo Resource*	*RelationTo Resource*	*RelationTo Resource*	*RelationTo Resource*				

ZDB		86.261	97.922	10.018	392.555	444.440	497.262				1.528.458
											alle Referenzen zu Norm-, Titeldaten (ohne Referenzen, die von ZDB-Dubletten ausgehen)
DNB ohne ZDB-Dubletten*	19.384.733										31.536.618
	archdesc	*persname*	*corpname*	*geogname*	*subject*	*level*					alle Referenzen
	ResourceName	*RelationToPerName*	*RelationToCorpName*	*RelationToGeoName*	*RelationToTopicTerm*	*RelationToResource*					
KPE	26.752	6.007.129	850.794	1.151.174	174.788	4.359.421					12.543.306

* Der Datendump der DNB wurde in insgesamt vier Teilen übermittelt. Die Ausdifferenzierung dieser Datensätze hat keinen weitergehenden Nutzen für den vorliegenden Bericht, daher ist hier lediglich die Gesamtauszählung aufgeführt.

M3 Anzahl der gültigen Referenzen je Datenfeld (gemäß Datenmodell für Relation) und Datendump

RelationTo	**PerName**	**CorpName**	**MeetName**	**UniTitle**	**TopicTerm**	**GeoName**	**ChronTerm**	**IsilTerm**	**Resource**	**alle Referenzen**
Felder	*500*	*510*	*511*	*530*	*550*	*551*	*548*	*Präfix 035*	*77X, 78X*	
Personen	347.542	669.514	201	2.280	4.053.424	1.825.148	4.280.076	10.175.320	0	21.353.505
Körperschaften	13.732	599.392	433	2.322	300.628	1.105.113	257.406	2.975.422	0	5.254.448
Kongresse	903	36.363	40.538	12	32.344	642.751	672.473	1.628.088	0	3.053.472
Werke	380.672	26.457	311	35.122	63.605	24.907	182.039	770.600	0	1.483.713
Sachbegriffe	2.793	3.539	18	133	211.268	12.020	7.314	424.270	0	661.355
Geografikum	8.966	3.338	24	27	115.639	178.214	47.533	616.394	0	970.135
GND	754.608	1.338.603	41.525	39.896	4.776.908	3.788.153	5.446.841	16.590.094	0	32.776.628
DNB ohne ZDB-Dubletten*	8.035.213	1.861.711	176.486	14.099	0	64	0	31.782.859	1.548.907	43.419.339
ZDB	16.610	1.057.662	45.169	3	0	201.168	0	2.856.175	1.479.072	5.655.859
Felder	*persname*	*corpname*			*subject*	*geogname*		*@source = 'ISIL'*	*level*	
KPE	5.824.034	841.214	0	0	174.709	1.151.171	0	4.327.785	4.359.421	16.678.334
Normdaten	754.608	1.338.603	41.525	39.896	4.776.908	3.788.153	5.446.841	16.590.094	0	**32.776.628**
Titeldaten	13.875.857	3.760.587	221.655	14.102	174.709	1.352.403	0	38.966.819	7.387.400	**65.753.532**
alle Daten	14.630.465	5.099.190	263.180	53.998	4.951.617	5.140.556	5.446.841	55.556.913	7.387.400	**98.530.160**

* Der Datendump der DNB wurde in insgesamt vier Teilen übermittelt. Die Ausdifferenzierung dieser Datensätze hat keinen weitergehenden Nutzen für den vorliegenden Bericht, daher ist hier lediglich die Gesamtauszählung aufgeführt.

M4 Anzahl der gemäß M3 referenzierten, gültigen Entitäten je Entitätentyp und Datendump

	PerName	**CorpName**	**MeetName**	**UniTitle**	**TopicTerm**	**GeoName**	**ChronTerm**	**IsilTerm**	**Resource**	**alle Referenzen**
Felder	*500*	*510*	*511*	*530*	*550*	*551*	*548*	*Präfix 035*	*77X, 78X*	
Personen	12.505	265.434	318	2.296	9.317	60.692	50.771	2	0	401.335
Körperschaften	5.310	2.828	19	26	2.850	71.353	11.755	2	0	94.143
Kongresse	785	16.463	32.623	12	1.420	20.929	33.622	2	0	105.856
Werke	167.985	123.295	142	2.076	14.696	71.354	468.728	2	0	848.278
Sachbegriffe	1.679	1.484	15	85	50.887	4.261	3.127	2	0	61.540
Geografikum	87.689	7.836	184	2.270	5.613	1.849	13.533	2	0	118.976
DNB1 ohne ZDB-Dubletten	532.527	94.653	8.375	222	0	1	0	2	344.214	979.994
DNB2 ohne ZDB-Dubletten	624.286	131.374	90.716	144	0	2	0	6	334.580	1.181.108
DNB3 ohne ZDB-Dubletten	538.724	76.051	37.323	2.773	0	1	0	6	545.850	1.200.728
DNB4 ohne ZDB-Dubletten	261.215	52.143	9.142	6.523	0	46	0	7	283.309	612.385
GND	256.515	347.857	33.143	6.598	60.994	127.253	537.054	2	0	1.369.416
DNB ohne ZDB-Dubletten	1.443.639	281.644	139.829	9.181	0	50	0	7	1.499.266	3.373.616
ZDB	8.819	387.313	34.690	3	0	20.864	0	2	911.248	1.362.939
Felder	*persname*	*corpname*			*subject*	*geogname*		*@source = 'ISIL'*	*level*	
KPE	284.544	45.486	0	0	5.811	6.157	0	602	4.326.296	4.668.896
Normdaten	256.515	347.857	33.143	6.598	60.994	127.253	537.054	2	0	**1.369.416**
Titeldaten	1.618.500	618.029	173.319	9.184	5.811	24.079	0	609	6.730.526	**9.180.057**
alle Daten	1.759.713	741.388	190.412	15.673	63.429	129.411	537.054	611	6730526	**10.168.217**

M5 Anzahl der als fehlerhaft identifizierten Datensätze je Datendump

Datendump	ohne gültige Identifikatoren	fehlende Normdaten (Tn)	fehlende Titeldaten
GND	1.929.635	37	0
DNB (ohne ZDB-Dubletten)	9.677.777	10.099.596	30.475
ZDB	52.766	5.520	70
KPE	55	192.702	0

In der KPE 1.447.006 Relationen gefunden, die nicht zu Entitäten aus der GND waren (z.B. @source='SLA')

M6 Anzahl der Dubletten je Datendump auf Ebene eines Datensatzes (zu DNB übertragene ZDB-Datensätze).

Datenfile	Anzahl
DNB*	541.840

* Der Datendump der DNB wurde in insgesamt vier Teilen übermittelt. Die Ausdifferenzierung dieser Datensätze hat keinen weitergehenden Nutzen für den vorliegenden Bericht, daher ist hier lediglich die Gesamtauszählung aufgeführt.

M7 Anzahl der Dubletten (nach Datenmodell) je Datendump auf Ebene eines Datenfeldes/Unterfeldes.

Datendump /Felder										
	500	*510*	*511*	*530*	*550*	*551*	*548*			**5XX**
	RelationToPerName	*RelationToCorpName*	*RelationToMeetName*	*RelationToUniTitle*	*RelationToTopicTerm*	*RelationToGeoName*	*RelationToChronTerm*			
Personen	56	128	0	1	879	423	54			1.541
Körperschaften	3	47	0	0	15	42	7			114
Kongresse	0	7	4	0	4	14	5			34
Werke	18	0	0	2	10	0	66			96
Sachbegriffe	0	0	0	0	4	1	1			6
Geografika	0	0	0	0	1	7	1			9
GND	77	182	4	3	913	487	134			1.800
	100	*700*	*110*	*710*	*111*	*711*	*130*	*730*	*751*	**1XX/7XX**
	RelationToPerName	*RelationToPerName*	*RelationToCorpName*	*RelationToCorpName*	*RelationToMeetName*	*RelationToMeetName*	*RelationToUniTitle*	*RelationToUniTitle*	*RelationToGeoName*	
ZDB	0	0	1	20	0	0	0	0	83	104
	770	*772*	*775*	*776*	*780*	*785*				**7XX**

	RelationToResource	*RelationToResource*	*RelationToResource*	*RelationToResource*	*RelationToResource*	*RelationToResource*				
ZDB	30	27	3	74	79	564				777
										alle Referenzen
DNB1 ohne ZDB-Dubletten										369
DNB2 ohne ZDB-Dubletten										3.246
DNB3 ohne ZDB-Dubletten										49.225
DNB4 ohne ZDB-Dubletten										39.450
DNB ohne ZDB-Dubletten										92.290

N1 Anzahl der Kanten gesamt

34.511.952

N2 Anzahl der Knoten je Datendump

Datendump	Anzahl
GND	8.295.047
DNB	19.384.733
ZDB	1.908.334
KPE	4.386.173
ChronTerm GND	537.054
IsilTerm	611

N3 Anzahl der Knoten je Entitätentyp

Entitaetentyp	Anzahl
CorpName	1.487.711
GeoName	308.197
MeetName	814.044
PerName	5.087.660
TopicTerm	212.135
UniTitle	385.300
ChronTerm	537.054
IsilTerm	611
Resource	25.679.240

N4 Anzahl der Knoten je Relationentyp gemäß Datenmodell und Datendump

Entitaetentyp*	PerName	CorpName	MeetName	UniTitle	TopicTerm	GeoName	ChronTerm	Resource	alle Entitaeten
GND	5.087.660	1.487.711	814.044	385.300	212.135	308.197	537.054	0	8.832.101
DNB ohne ZDB-Dubletten	0	0	0	0	0	0	0	19.384.733	19.384.733
ZDB	0	0	0	0	0	0	0	1.908.334	1.908.334
KPE	0	0	0	0	0	0	0	4.386.173	4.386.173
gesamt	5.087.660	1.487.711	814.044	385.300	212.135	308.197	537.054	25.679.240	34.511.341

*Eine Ausdifferenzierung der Knoten des Entitätentyps IsilTerm (N=611) nach Datendump ist hier aufgrund der Auszählung auf Datenfeldebene nicht möglich. Die Gesamtzahl der Entitäten unterscheidet sich daher von N1 um die Anzahl der IsilTerms.

N5 Anzahl der Kanten gesamt

98.530.160

N6 Anzahl der Kanten je Datendump

Datendump	Anzahl (ohne RelationToIsilTerm)	Anzahl (mit RelationToIsilTerm)
GND	16.186.534	32.776.628
ZDB	2.799.684	5.655.859
DNB	11.636.480	43.419.339
KPE	12.350.549	16.678.334

N7 Anzahl der Kanten je Relationentyp gemäß Datenmodell

Relationentyp	Anzahl
RelationToPerName	14.630.465
RelationToCorpName	5.099.190
RelationToMeetName	263.180
RelationToUniTitle	53.998
RelationToTopicTerm	4.951.617
RelationToGeoName	5.140.556
RelationToChronTerm	5.446.841
RelationToIsil	55.556.913
RelationToResource	7.387.400

N8 Anzahl der Kanten je Relationentyp gemäß Datenmodell und Datendump

RelationTo	PerName	CorpName	MeetName	UniTitle	TopicTerm	GeoName	ChronTerm	IsilTerm	Resource
GND	754.608	1.338.603	41.525	39.896	4.776.908	3.788.153	5.446.841	16.590.094	0
DNB	8.035.213	1.861.711	176.486	14.099	0	64	0	31.782.859	1.548.907
ZDB	16.610	1.057.662	45.169	3	0	201.168	0	2.856.175	1.479.072
KPE	5.824.034	841.214	0	0	174.709	1.151.171	0	4.327.785	4.359.421