Ehrenfeld, Wilfried

Research Report

Research Explorer – Technische Dokumentation der Routinen

IWH Technical Reports, No. 03/2015

Provided in Cooperation with:
Halle Institute for Economic Research (IWH) – Member of the Leibniz Association

Suggested Citation: Ehrenfeld, Wilfried (2015) : Research Explorer – Technische Dokumentation der Routinen, IWH Technical Reports, No. 03/2015, Leibniz-Institut für Wirtschaftsforschung Halle (IWH), Halle (Saale)

This Version is available at:
https://hdl.handle.net/10419/144720

Leibniz-Institut für
Wirtschaftsforschung
Halle

IWH TECHNICAL REPORTS

Research Explorer

Technische Dokumentation der Routinen

Wilfried Ehrenfeld

03|2015

Autor:
Dr. Wilfried Ehrenfeld

Kontakt:
Dr. Cornelia Lang
Leiterin des IWH-Datenzentrums
Telefon: + 49 345 77 53 802
Fax: + 49 345 77 53 820
E-Mail: cornelia.lang@iwh-halle.de

Herausgeber: LEIBNIZ-INSTITUT FÜR WIRTSCHAFTSFORSCHUNG HALLE – IWH
Geschäftsführender Vorstand: Prof. Reint E. Gropp, Ph.D.
Prof. Dr. Oliver Holtemöller
Dr. Tankred Schuhmann

Hausanschrift: Kleine Märkerstraße 8, D-06108 Halle (Saale)
Postanschrift: Postfach 11 03 61, D-06017 Halle (Saale)
Telefon: +49 345 7753 60
Telefax: +49 345 7753 820
Internetadresse: www.iwh-halle.de

Zitierhinweis:
Ehrenfeld, Wilfried: Research Explorer – Technische Dokumentation der Routinen. IWH Technical Reports 03/2015. Halle (Saale) 2015.

ISSN 2365-9076

Research Explorer

Technische Dokumentation der Routinen

Zusammenfassung

Der Research Explorer ist ein von der Deutschen Forschungsgemeinschaft (DFG) und der Deutschen Akademischen Austauschdienstes (DAAD) zur Verfügung gestelltes Verzeichnis von Instituten an deutschen Hochschulen sowie außeruniversitäre Forschungseinrichtungen. Er steht kostenlos zum Download bereit und umfasst ca. 23.000 Einträge. Uns liegen zwei verschiedene Abzüge dieses Verzeichnisses vor, die sich jedoch in Abdeckung und Tiefe unterscheiden. Die eine Version umfasst 1712 Einträge, die andere über 22.000.

Ziel der hier vorgestellten Routinen ist der Abgleich und die Ergänzung dieser zwei Abzüge des Research Explorers. Die Notwendigkeit ergibt sich aus der unterschiedlichen Tiefe und Abdeckung dieser beiden Versionen. Weiter sind beide Versionen an einigen Stellen unvollständig, so dass einzelne Institutionen nachgetragen werden mussten. Die resultierenden Datensätze des modifizierten Research Explorers können anschließend etwa zur Abbildung von Kooperationsbeziehungen genutzt werden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Ausgangsbasis und Umriss des Verfahrens

Der Research Explorer (kurz: REX) ist ein Verzeichnis von Instituten an deutschen Hochschulen sowie außeruniversitären Forschungseinrichtungen. Er ist das gemeinsame Produkt der Deutschen Forschungsgemeinschaft (DFG) und des Deutschen Akademischen Austauschdienstes (DAAD) in Zusammenarbeit mit der Hochschulrektorenkonferenz (HRK). Dieses Verzeichnis wird als Online-Datenbank[1] kostenlos bereitgestellt und umfasst knapp 23.000 Einträge. Die Angaben sind nach geografischen, fachlichen und strukturellen Kriterien geordnet und umfassen neben dem Namen der Institution Angaben zum Standort (Variablen Strasse; Hausnummer; PLZ; Ortsname; Bundesland) und Typ (Variablen Fachgebiet; Einrichtungstyp; Sektion).

Uns liegen zwei Abzüge dieses Verzeichnisses vor, die sich in Abdeckung und Tiefe unterscheiden. Dabei gilt es eine „kurze“ oder „kleine“ Version und eine „lange“ oder „große“ Version des REX zu unterscheiden. Die „kurze“ Version umfasst 1 712 Einträge (Stand: 21.03.2014). Die „lange“ Version beinhaltet 22 331 Einträge (Stand: 07.05.2014). In letzterer ist für die Universitäten die Untergliederung bis auf Lehrstuhlebene abgebildet, nicht hingegen in der kurzen Version, in der nur die Hochschule als übergeordnete Einheit gelistet ist. Weiter umfasst die lange Version auch die Universitätskliniken, die in der kurzen Version gänzlich fehlen.

In beiden Versionen des Research Explorers wurden die Identifikationsnummern für Lehrstühle, Standorte von Forschungseinrichtungen und ähnliche Strukturen nicht hierarchisch vergeben, sondern voneinander unabhängig. Für die Forschungsarbeit wäre hingegen ein hierarchisches System von IDs wünschenswert gewesen. Auch unterscheiden sich an mehreren Stellen die Schreibweisen der Institutionen zwischen der kurzen und der langen Version. Außerdem sind beide Versionen an einigen Stellen unvollständig, so dass einzelne Institutionen nachgetragen werden mussten. Dies betrifft vor allem die Standorte der Fraunhofer-Gesellschaft und anderer außeruniversitärer Forschungseinrichtungen.

Ziel der hier vorgestellten Routinen ist daher der Abgleich und die Ergänzung dieser zwei Abzüge des Research Explorers. Die Notwendigkeit des Abgleichs ergibt sich aus der unterschiedlichen Tiefe und Abdeckung der beiden Versionen. Dabei sollten nicht einfach nur die unterschiedlichen Einträge der Datenbanken verglichen werden, sondern es sollte die kurze Version um ein verkürztes hierarchisches System von Standorten aus der langen Version ergänzt werden. Als Aggregationsmerkmal wurde dabei die Kreiskennziffer gewählt.

Zur Durchführung wurde daher ein Filter entwickelt, der ausgehend von einer Erweiterung der kurzen Version des REX alle Einträge der langen Version auf Zugehörigkeit der Institution zu einer Institution der erweiterten kurzen Version überprüft. Die resultierenden Datensätze des modifizierten Research Explorers können anschließend - etwa in Verbindung mit Daten aus der Unternehmensdatenbank Amadeus - unter Verwendung von Record-Linkage-Techniken

[1] http://www.research-explorer.de

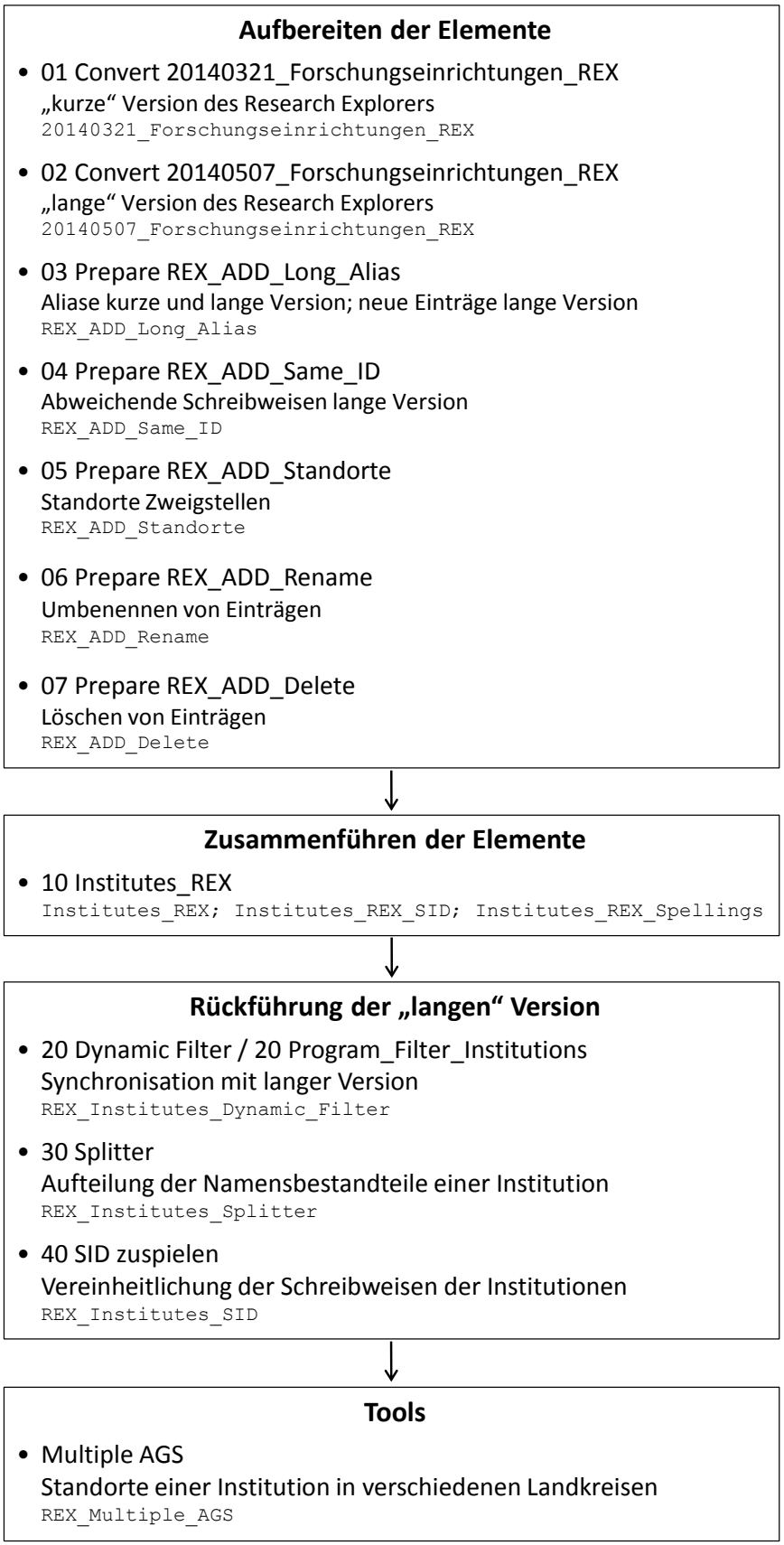

Abbildung 1: Ablauf und resultierende Datensätze des verwendeten Verfahrens.

(siehe Ehrenfeld 2015c) zur Abbildung von Kooperationsbeziehungen genutzt werden (siehe hierzu Titze et al. 2015 und Ehrenfeld 2015a).

Abbildung 1 zeigt den strukturellen Ablauf der Aufbereitung. Im Folgenden werden die einzelnen Stufen des Verfahrens genauer beschrieben.

2. Aufbereitung der einzelnen Bestandteile

In diesem Schritt werden die einzelnen Bestandteile des erweiterten REX-Datensatzes eingelesen und aufbereitet.

01 Convert 20140321_Forschungseinrichtungen_REX.do

Die Daten der kleinen Version des Research Explorer werden aus der xlsx-Datei eingelesen. Das vorhandene Id-Feld wird in ein Format mit konstanter Länge konvertiert (REXid). Die ID umfasst hier 12 Stellen, wobei die ersten 3 Zeichen die Herkunft (REX) identifizieren. Danach folgen 9 Stellen, die numerisch identisch mit der Nummer aus dem ursprünglichen Id-Feld sind.

Die Felder werden um störende Steuerzeichen (`CR`; `LF`; `TAB`; `QUOTES`) sowie überflüssige (d. h. am Anfang oder Ende stehende) und doppelte Leerzeichen bereinigt. Einrichtungstyp und Sektion werden von Strings in numerische Werte konvertiert und mit Labels versehen (siehe Anhang A.2 und A.3). Die einzelnen Variablen werden bereinigt, formatiert (siehe Anhang A.1) und sortiert (Sortierreihenfolge: Institution - PLZ - Ortsname).

Input: `20140321_Forschungseinrichtungen_REX.xlsx`
Id; Institution; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

Output: `20140321_Forschungseinrichtungen_REX.dta`
REXid; Institution; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

Output: `20140321_Forschungseinrichtungen_REX.tsv`
REXid; Institution; PLZ; Ortsname

02 Convert 20140507_Forschungseinrichtungen_REX.do

Die Routine leistet dasselbe wie 01 Convert 20140321_Forschungseinrichtungen_REX.do für den großen Research Explorer Datensatz. Zusätzlich wird eine Version angelegt, bei der mittels einer externen Routine (`PLZ_AGS_Prog`) der Kreisschlüssel (AGS5) aus der PLZ und dem Ortsnamen abgeleitet wird.

In der langen Version gibt es neben der Postanschrift noch das Feld Institution. Im Gegensatz zum Feld gleichen Namens in der kurzen Version enthält dieses Feld Angaben zu strukturell untergeordneten Organisationselementen wie z. B. Lehrstühlen.

Input: `20140507_Forschungseinrichtungen_REX.xlsx`
Id; Institution; Postanschrift; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

Output: `20140507_Forschungseinrichtungen_REX.dta`
REXid; Postanschrift; Institution; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

Output: `20140507_Forschungseinrichtungen_REX_AGS.dta`
REXid; Postanschrift; Institution; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; *AGS5*; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

03 Prepare REX_ADD_Long_Alias.do

In der Tabelle `REX_ADD_Long_Alias.xlsx` sind drei Arten von Zusätzen erfasst:

- Alias-Schreibweisen von Institutionen aus der kurzen Version [`Status = 1`]
- Neue, aus der langen Version übernommene, Einträge [`Status = 2`]
- Alias-Schreibweisen von Institutionen aus der langen Version [`Status = 3`]

Die Routine liest die Excel-Datei ein und bereitet diese auf (Generierung REXid; Festlegung von Variablentypen; Bereinigung von Leerzeichen). Im Zuge dessen erfolgt auch die Großschreibung von Institutionsnamen (Originalschreibweise; Postanschrift; Institution) und Ortsname (Externe Routine `WEupper`). Der Name der Institution wird dabei in Institution_add umbenannt. Anschließend werden anhand der REXid fehlende Daten aus der kurzen Version (`20140321_Forschungseinrichtungen_REX.dta`) bzw. der langen Version (`20140507_Forschungseinrichtungen_REX.dta`) hinzugefügt.

Zuerst wird der Datensatz anhand von REXid um Angaben aus der kurzen Version ergänzt. Anhand der Variable `_merge` wird bestimmt, dass es sich bei den hinzugefügten Daten um Daten aus der kurzen Version handelt. Die Variable Status gibt danach die Quelle der Ergänzungen an. Für die so erkannten und ergänzten Datensätze gilt: `Status = 1`. Die noch nicht zugeordneten Datensätze haben `Status = 0`.

Durch Vergleich des Feldes Originalschreibweise mit Institution aus der kurzen Version wird anschließend festgestellt, ob es sich bei dem Eintrag um einen validen Alias für einen bereits bestehenden Eintrag handelt. Beim Abgleich der Aliase mit der Institution der kurzen

Version gibt die Variable test das Ergebnis der Prüfung an. Diese Variable kann dabei drei Zustände annehmen:

- Keine Originalschreibweise vorhanden [`test = 0`]
- Die Originalschreibweise ist vorhanden, aber abweichend [`test = 1`].
  Sie sollte daher korrigiert werden
- Die Originalschreibweise ist OK [`test = 2`]

Anschließend wird der Datensatz mit der langen Version verknüpft. Die Identifikation von neuen, aus der langen Version übernommenen Einträgen erfolgt anhand des Namens der Institution (Institution_add) mit dem Feld Postanschrift, unter der Bedingung, dass Originalschreibweise leer ist. Die so identifizierten Datensätze erhalten `Status = 2`. Auch hier gibt die Variable test das Ergebnis bzgl. der Originalschreibweise an. Für die neu aus der langen Version hinzugefügten Datensätze muss Originalschreibweise leer sein (`test = 2`).

Die Aliase der langen Version erhalten schließlich `Status = 3`. In Analogie zu den Aliasen für die kurze Version gibt auch hier die Variable test das Ergebnis des Abgleichs der Schreibweisen an. Dazu darf die Institution (Institution_add) nicht dem Feld Postanschrift entsprechen.

Abschließend wird geprüft, ob jeder Alias der langen Version auch einen entsprechenden Eintrag in der langen Version besitzt. Hierzu werden root und root2 verwendet. Die Variable root erhält den Wert 1, wenn der Eintrag ein Alias aus der langen Version ist (`Status = 3`), sonst 0. Variable root2 ist anschließend definiert als das Minimum von root für jede (Gruppe von) REXid. In jeder Gruppe von REXid müssen also auch die neu hinzugefügten Einträge aus der langen Version vorhanden sein (`Status = 2`). Für diese gilt jedoch: `root = 0`. Wenn das Minimum von root2 also 0 ist, besitzt der Alias in der REXid-Gruppe einen entsprechenden Stamm-Eintrag aus der langen Version. Wenn `root2 = 1`, dann fehlt ein solcher Eintrag. Dies muss dann korrigiert werden.

Das Ergebnis dieser Routine sind drei Datensätze:

- ein Datensatz, um die Ergebnisse dieser Routine zu evaluieren. In diesem sind auch alle Aliase und übernommenen Einträge aus der langen Version enthalten [`REX_ADD_Long_Alias_Eval.dta`].
- ein Datensatz mit allen Aliasen und den neuen, aus der langen Version übernommenen Einträgen. In diesem Datensatz wird die Variable Status umbenannt (source) und umgeformt, so dass sie auch die Quelle des ergänzenden Datensatzes erfasst (`Status = 1` $\rightarrow$ `source = 31` - siehe Anhang A.4) [`REX_ADD_Long_Alias.dta`].
- ein SID-Datensatz[2], der nur die neu hinzugefügten Daten aus der langen Version (`Status = 2`) enthält [`REX_ADD_Long_Alias_SID.dta`].

[2] SID bedeutet: **S**ingle **ID**. Es existiert in diesem Datensatz also genau ein Eintrag für jede ID.

Input: `REX_ADD_Long_Alias.xlsx`
Id; Institution; Originalschreibweise; Kommentar

Output: `REX_ADD_Long_Alias_Eval.dta`
REXid; Status; test; Institution_add; Postanschrift; Originalschreibweise; Institution; Kommentar; Strasse; Hausnummer; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; PLZ; Ortsname

Output: `REX_ADD_Long_Alias.dta`
REXid; Institution; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

Output: `REX_ADD_Long_Alias_SID.dta`
REXid; Status; test; Institution_add; Postanschrift; Originalschreibweise; Institution; Kommentar; Strasse; Hausnummer; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; PLZ; Ortsname

04 Prepare REX_ADD_Same_ID.do

Diese Routine erkennt - gegeben derselben REXid - unterschiedliche Schreibweisen für Institution in der langen (Suffix: _long) und kurzen Version (Suffix: _short) des REX. Institutionsnamen und Ortsnamen werden in Großschreibung umgewandelt (Externe Routine `WEupper`). Die Identifikation erfolgt mittels eines Scoring-Systems. In diese fließen der Institutionsname (`score = 4`), PLZ (`score = 2`) und Ortsname (`score = 1`) ein. Wenn mindestens eine dieser Angaben abweichend ist, wird der Eintrag in die Ergänzungstabelle aufgenommen (die Überprüfung zeigt jedoch, dass hier `score` entweder 0 oder 7 ist). Die Daten werden isoliert, aufbereitet, sortiert und mit einer Quellenangabe versehen (`source = 2`; siehe Anhang A.4).

Die Datei `REX_ADD_Same_ID_Eval` enthält im Vergleich zu `REX_ADD_Same_ID` zusätzliche Variablen, die zur Evaluierung der Routine dienen.

Input: `20140507_Forschungseinrichtungen_REX.dta`
`20140321_Forschungseinrichtungen_REX.dta`

Output: `REX_ADD_Same_ID_Eval.dta`
REXid; Institution_long; Institution_short; PLZ_long; PLZ_short; Ortsname_long; Ortsname_short; Strasse; Hausnummer; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

Output: `REX_ADD_Same_ID.dta`
REXid; Institution; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

05 Prepare REX_ADD_Standorte.do

Die Datei `REX_ADD_Standorte.xlsx` enthält Standorte von bereits im REX existierenden Institutionen und deren verschiedene Schreibweisen. Hierzu wurde ein hierarchisches ID-System auf Basis der (bereits existierenden) IDs der „Haupthäuser“ aufgebaut. Jede (meist untergeordnete) Zweigstelle bekommt dabei die ID des „Haupthauses“ zugeordnet und zusätzlich einen mit „-“ abgetrennten Suffix, welcher dreistellig die laufende Nummer der Zweigstelle angibt. Diese REXid-Felder sind folglich 16 Stellen lang.

Nachdem die Daten aus der Excel-Tabelle eingelesen wurden, erfolgt die Umwandlung in Großschreibung von Institution (Externe Routine `WEupper`). Die Variablen werden aufbereitet, Duplikate werden gelöscht und Einrichtungstypen zugeordnet (externe Routinen `WEreplace_Einrichtungstyp`; `RDlabel_Einrichtungstypen`; `RDlabel_Sektionen` - siehe Anhänge A.2 und A.3). Benötigte Daten werden aus der kurzen Version ergänzt. Strasse und Ortsname werden ebenfalls in Großschreibung umgewandelt.

Anschließend wird die Variable Status erzeugt, welche angibt, ob es sich bei einem Eintrag um einen Alias für einen Standort handelt. Ist der Eintrag ein Alias eines existierenden Eintrages, so gilt: `Status = 1`, sonst 0. Die Überprüfung erfolgt anhand des Inhaltes von Originalschreibweise. Ist dieses Feld nicht leer, so handelt es sich um einen Alias.

Schließlich werden die Daten mit einer aus Status abgeleiteten Quellenangabe versehen (`source = 40` bzw. `source = 41`; siehe Anhang A.4). Die Tabelle `REX_ADD_Standorte_Eval.dta` enthält zusätzliche Informationen zur Evaluierung der Routine. Für die SID-Tabelle `REX_ADD_Standorte_SID.dta` werden alle Aliase gelöscht.

Input: `REX_ADD_Standorte.xlsx`
REXid_ext; Id-Bezug; Institution; Originalschreibweise; Strasse; Hausnummer; PLZ; Ortsname; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; Kommentar

Output: `REX_ADD_Standorte_SID.dta`
REXid_ext; REXid; Status; Institution; Originalschreibweise; Strasse; Hausnummer; PLZ; Ortsname; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; Kommentar; OrtsnamemitZusatz; Bundesland

Output: `REX_ADD_Standorte_Eval.dta`
REXid_ext; REXid; Status; Institution; Originalschreibweise; Strasse; Hausnummer; PLZ; Ortsname; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; Kommentar; OrtsnamemitZusatz; Bundesland

Output: `REX_ADD_Standorte.dta`
REXid; Institution; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

06 Prepare REX_ADD_Rename.do

In `REX_ADD_Rename.xlsx` sind zu ändernden Schreibweisen der kurzen REX-Version enthalten. Dies ist notwendig, wenn die dort vorgegebene Schreibweise nicht dem aktuellen, tatsächlichen Stand entspricht. Die „falschen“ Schreibweisen sollen dennoch als Aliase erhalten bleiben, um diese Einträge beim Vergleich mit der langen Version erkennen zu können.

Die Daten werden aus der Excel-Tabelle eingelesen. Originalschreibweise und Institution werden in Großschreibweise umgewandelt. Institution wird in Institution_ren umbenannt, um Daten aus der kurzen Version ergänzen zu können. Die Daten werden ergänzt und der korrekte Bezug eines Alias zur Institution wird überprüft.

Entspricht die Originalschreibweise dem Feld Institution der kurzen Version(!), so ist der Alias valide und es gilt `test = 1`, sonst 0. Weiter wird geprüft, ob das Institution-Feld der kurzen Version dem Institution-Feld der Rename-Tabelle (Institution_ren) entspricht (Variable test2). Ist dies der Fall (test2 = 1), so ist die Aufnahme dieser Ersetzung zwecklos, da ja nichts umbenannt werden muss.

Die Eval-Tabelle `REX_ADD_Rename_Eval.dta` enthält zusätzliche Informationen zur Evaluierung der Routine.

Input: `REX_ADD_Rename.xlsx`
Id; Institution; Originalschreibweise; Kommentar

Output: `REX_ADD_Rename_Eval.dta`
REXid; Institution_ren; Originalschreibweise; Kommentar; Institution; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

Output: `REX_ADD_Rename.dta`
REXid; Institution; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

07 Prepare REX_ADD_Delete.do

Die Tabelle `REX_ADD_Delete.xlsx` stellt eine „Droplist“ zur Verfügung. Alle hier verzeichneten Einträge sollen später aus dem Datensatz gelöscht werden. Die Notwendigkeit eines solchen Vorgehens ergibt sich aus zwei Sachverhalten. Zum einen ist die Aufnahme von Standorten von Forschungsinstituten im Research Explorer unvollständig (vgl. hierzu auch Abschnitte 05 Prepare REX_ADD_Standorte.do und Multiple AGS.do). Zum anderen wurden Einträge unter leicht abweichender Schreibweise (oder deutsch/englisch) mehrfach aufgenommen.

Die Daten werden aus der Excel-Tabelle `REX_ADD_Delete.xlsx` eingelesen. Die 9-stellige REXid wird aus dem ID-Feld generiert. Institution wird in Großschreibung umgewandelt. Die Daten werden bereinigt und formatiert.

Input: `REX_ADD_Delete.xlsx`
Id; Replace_ID; Institution; Kommentar

Output: `REX_ADD_Delete.dta`
REXid; Replace_ID; Institution; Kommentar

3. Zusammenfügung der Datensätze und Abgleich mit der langen Version

In diesem Schritt werden die aufbereiteten Bestandteile des Datensatzes zusammengefügt und finalisiert.

10 Institutes_REX.do

Hier werden die bisher erstellten und extrahierten Datensätze zusammengefügt. Im Zuge dessen werden drei Gruppen von Datensätzen erstellt, die später benötigt werden:

- `Institutes_REX.tsv` ist der resultierende Suchfilter, mit dem anschließend (siehe 20 Dynamic Filter.do) die lange Version des REX durchsucht wird. Die `dta`-Version entspricht inhaltlich der `tsv`-Version.
- **`Institutes_REX_Eval.dta`** kann als das erste Hauptergebnis des Verfahrens angesehen werden. Sie enthält alle REXids mit sämtlichen Schreibweisen und stellt so die erweiterte (kurze) Version des REX als Identifizierungstabelle dar.
- **`Institutes_REX_SID.dta`** ist das zweite Hauptergebnis des Verfahrens. Ebenso wie `Institutes_REX_Eval.dta` enthält sie alle REXid-Einträge, jedoch mit dem Unterschied, dass für jeden REXid-Eintrag nur eine Schreibweise erfasst ist. Sie ist also eine SID-Liste mit eindeutigen REXid-Nummern und wird später zur Standardisierung von identifizierten Einträgen verwendet (siehe 40 SID zuspielen.do).
- `Institutes_REX_SID_suffix.dta` ist dieselbe Liste, jedoch sind die Felder mit dem Suffix `_SID` versehen.
- `Institutes_REX_Spellings.dta` enthält die verschiedenen Schreibweisen von Institution für eine REXid. Sie entspricht von der Abdeckung her der Tabelle `Institutes_REX.dta`, enthält aber nur die REXid und die Institution. Sie wird in Multiple AGS.do verwendet, um alternative Schreibweisen von Standorten in mehreren Landkreisen zu erzeugen.

Der Datensatz wird mittels `append` aus den in Abschnitt 2 eingelesenen und aufbereiteten Datensätzen zusammengefügt. Zusätzlich werden die in Multiple AGS.do erzeugten Datensätze von Institutionen mit mehreren Standorten in verschiedenen Landkreisen zugespielt. Diese Daten erhalten die Quellencodierung `source = 50` bzw. `source = 51` (siehe Anhang A.4).

Anmerkung: *Soll die in Multiple AGS.do erstellte Tabelle REX_Multiple_AGS.dta neu erzeugt werden, so muss die Einfügung dieses Datensatzes (append) hier deaktiviert und der Rest der Routine neu ausgeführt werden. Multiple AGS.do erzeugt anschließend eine neue REX_Multiple_AGS.dta.*

Die Daten aus der kurzen Version des REX (20140321_Forschungseinrichtungen_REX.dta) erhalten die Quellencodierung source = 1 (siehe Anhang A.4). Die Daten aus der Rename-Datei (REX_ADD_Rename.dta) erhalten die Quellencodierung source = 0. Die Codierung der Quellen in der Variable source entspricht Anhang A.4.

Die Delete-Tabelle (REX_ADD_Delete.dta) wird angefügt (merge m:1 REXid) und die so identifizierten Datensätze gelöscht. Die Daten werden bereinigt, Institution, Ortsname, und Strasse werden in Großschreibung überführt. Der Datensatz wird nach den Variablen Institution, REXid, PLZ und Ortsname sortiert und Duplikate dieser Variablen werden identifiziert (dup), zur Löschung gekennzeichnet (drp) und schließlich gelöscht.

Abschließend wird die Kreiskennziffer (AGS5) aus der PLZ und dem Ortsnamen abgeleitet (externe Routine PLZ_AGS_Prog), die Variablen sortiert und die Länge des String-Feldes Institution bestimmt (len). Für die Filter-Tabelle Institutes_REX.tsv wird diese Liste dieser Länge nach absteigend geordnet. Dies hat zur Folge, dass bei der späteren Filterung Spezialfälle (z.B. Universität ... Klinikum) vor den allgemeineren Fällen (Universität ...) gefunden werden.

Für die SID-Tabelle (Institutes_REX_SID.dta) werden die Datensätze aus den Quellen mit den source-Codes 2 (Alias lange vs. kurze Version), 31 (Alias kurze Version), 33 (Alias lange Version), 41 (Alias Standort) und 51 (Alias Multiple AGS) entfernt (siehe Anhang A.4). Im Zuge dessen wird auch Code 32 zu Code 3 umgewandelt. Duplikate bzgl. REXid werden gelöscht.

Input: REX_Multiple_AGS.dta
REX_ADD_Standorte.dta
REX_ADD_Long_Alias.dta
REX_ADD_Same_ID.dta
20140321_Forschungseinrichtungen_REX.dta
REX_ADD_Rename.dta
REX_ADD_Delete.dta

Output: Institutes_REX_Eval
REXid; Institution; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; AGS5; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source; len

Output: Institutes_REX.dta
REXid; Institution; PLZ; Ortsname; AGS5

Output: Institutes_REX.tsv
[REXid; Institution; PLZ; Ortsname; AGS5]

Output: `Institutes_REX_Spellings.dta`
REXid; Institution

Output: `Institutes_REX_SID.dta`
REXid; Institution; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; AGS5; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

Output: `Institutes_REX_SID_suffix.dta`
REXid_SID; Institution_SID; Strasse_SID; Hausnummer_SID; PLZ_SID; Ortsname_SID; OrtsnamemitZusatz_SID; AGS5_SID; Bundesland_SID; Internetadresse_SID; Fachgebiet_SID; Einrichtungstyp_SID; Sektion_SID; source_SID

20 Dynamic Filter.do

Die lange Version des Research Explorers (`20140507_Forschungseinrichtungen_REX.dta`) wird eingelesen. Der Datensatz wird aufbereitet, Postanschrift, Institution, Strasse und Ortsname werden in Großschreibung umgewandelt. Im Vorbereitung zur Filterung werden eine Reihe von Flags und temporären Variablen (nähere Erklärung in 20 Program_Filter_Institutions.do) angelegt (Tabelle 1).

Tabelle 1: Flags und temporäre Variablen in 20 Dynamic Filter.do

Name	Typ	Bedeutung
found	byte	Flag: Eintrag gefunden
level0	byte	Flag: Vollständige Entsprechung: `Inst_Short` = `Inst_Long`
subst	byte	Flag: Datensatz ersetzen?
score	byte	Flag: Übereinstimmung Datensätze Short Long
score_tmp	byte	Temp: Übereinstimmung Datensätze Short Long
REXid_Short	str12	REXid der kurzen Version
REXid_tmp	str12	Temp: REXid der kurzen Version
REXid_hist	str12	Temp: Ersetzungsgeschichte von REXid
Inst_Long	strL	Duplikat von Postanschrift
Inst_Short	strL	Institution der kurzen Version
Inst_tmp	strL	Temp: Institution der kurzen Version
PLZ_Short	str5	PLZ der kurzen Version
PLZ_temp	str5	Temp: PLZ der kurzen Version
Ortsname_Short	str24	Ortsname der kurzen Version
Ortsname_tmp	str24	Temp: Ortsname der kurzen Version
AGS_Short	str24	AGS5 der kurzen Version
AGS_tmp	str24	Temp: AGS5 der kurzen Version

Die Variablen werden formatiert und in die richtige Reihenfolge gebracht. Das Institutionsfeld aus der langen Version wird in Institution3 umbenannt. Die Filterroutine 20 Program_Filter_Institutions.do wird aufgerufen und die Verteilung von score angezeigt.

Dieser Filter dient dazu, den Einträgen aus der langen Version möglichst passgenaue Einträge aus der in 10 Institutes_REX.do erzeugten kurzen, ergänzten Version des REX (`Institutes_REX`) zuzuordnen. Die zugeordneten Daten stehen nach der Filterung in den Feldern mit dem Suffix _Short: REXid_Short; Inst_Short; PLZ_Short; Ortsname_Short und AGS_Short).

Der Variable level0 wird der Wert 1 bei Identität von `Inst_Short` = `Inst_Long` zugeordnet, sonst 0. Sie gibt an, dass es sich bei diesem Datensatz um die übergeordnete Institution selbst handelt, also Inst_Long keine weiteren Angaben zu untergeordneten Organisationselementen enthält. Die temporären Variablen (Suffix _tmp) werden abschließend gelöscht.

Input: `20140507_Forschungseinrichtungen_REX.dta`

Output: `REX_Institutes_Dynamic_Filter.dta`
found; level0; score; REXid; REXid_Short; Inst_Long; Inst_Short; Institution3; PLZ; PLZ_Short; Ortsname; Ortsname_Short; AGS5; AGS_Short; AGS_tmp; Postanschrift; Strasse; Hausnummer

20 Program_Filter_Institutions.do

Dieses Programm enthält den dynamischen Filter zum Abgleich der langen Version des REX mit der erweiterten kurzen Version. Die in 10 Institutes_REX.do erzeugte Filterdatei `Institutes_REX.tsv` wird geöffnet und die erste Zeile eingelesen. Nacheinander werden die durch das TAB-Zeichen getrennten Felder der tsv-Datei abgespalten und in lokale Variablen überführt: ID; Inst; PLZ; Ort und AGS.

Durch Vergleich der Variable Inst_Long (die ja ein Duplikat von Postanschrift ist) mit der lokalen Variablen Inst (aus der tsv-Datei) werden Kandidaten für die Zuordnung der Institutionen identifiziert. Wenn die lokale Variable Inst am Anfang von Inst_Long steht, könnte ein solcher Kandidat gefunden worden sein. In diesem Fall wird der Zähler der gefundenen Übereinstimmungen (Flag found) um eins erhöht und die Felder aus der tsv-Datei (ID; Inst; PLZ; Ort und AGS) in die temporären Variablen REXid_tmp; Inst_tmp; PLZ_tmp; Ortsname_tmp und AGS_tmp übernommen.

Anschließend wird mittels eine Scoring-Systems (siehe Tabelle 2) die Güte der Passung verschiedener Variablen ermittelt und der temporären Variable score_tmp zugeordnet.
Falls der Datensatz in den _tmp-Variablen nun einen höheren score (score_tmp) hat als der bisherige beste Kandidat in den _Short-Variablen (score), erhält die Flag subst den Wert 1 (sonst 0). Anschließend wird der _Short-Datensatz mit dem _tmp-Datensatz überschrieben und die bisherige REXid_Short der Variablen REXid_hist hinzugefügt.

Dann wird die nächste Zeile von Institutes_REX.tsv eingelesen. Dies geschieht so lange, bis keine weiteren Daten mehr in der tsv-Datei vorhanden sind. Als letztes wird die tsv-Datei geschlossen.

Input: `Institutes_REX.tsv`

Tabelle 2: Scores in 20 Program_Filter_Institutions.do

score	Bedeutung
32	Vollständige Gleichheit der Institutionsnamen Inst_Long und Inst_tmp.
16	Inst_tmp steht am Anfang von Inst_Long und ist länger als der bisherige Eintrag von Inst_Short. Diese Überprüfung dient dazu, erweiterte Einträge wie z.B. Universität ... Klinikum höher zu gewichten, als ihre übergeordneten und daher kürzeren Institutionen (Universität ...)
8	Inst_tmp steht am Anfang von Inst_Long
4	Übereinstimmung von PLZ und PLZ_tmp
2	Übereinstimmung von Ortsname und Ortsname_tmp
1	Übereinstimmung von AGS5 und AGS_tmp

30 Splitter.do

Diese Routine spaltet nach der Filterung die zusätzlichen Informationen aus dem Institutionsnamen der langen Version (Inst_Long) ab. Ziel des Splitters ist ein hierarchischer Aufbau der Institutionsebenen nach Institution1, Institution2 und Institution3.

In der langen Version des REX gibt es neben der Postanschrift (als Arbeitskopie in Inst_Long vorhanden) noch das Feld Institution. Im Gegensatz zum Feld gleichen Namens in der kurzen Version enthält dieses Feld Angaben zu strukturell untergeordneten Organisationselementen wie z. B. Lehrstühlen. Das Feld wurde in 20 Dynamic Filter.do in Institution3 umbenannt. Dieses Feld enthält die letzte untergeordnete Hierarchieebene.

Zu Beginn wird das Feld Institution3 geleert, falls sein Inhalt Inst_Long entspricht. Anschließend wird die in Inst_Short erkannte übergeordnete Institution aus Inst_Long entfernt, falls diese Angabe am Anfang steht. Der Rest von Inst_Long enthält dann die verbleibende mittlere und letzte Hierarchieebene. Folglich wird Inst_Short in Institution1 umbenannt und Inst_Long in Institution2.

Weiter wird Institution3 geleert, falls sein Inhalt mit Institution2 identisch ist. Falls der Inhalt von Institution3 vollständig am Ende von Institution2 steht, wird dieser Teil aus Institution2 entfernt. Schließlich wird Institution3 geleert, falls sein Inhalt in Institution2 vorhanden ist.

Die Einträge werden nach der REXid der kurzen Version (REXid_Short) gruppiert (Variable grp enthält die Gruppennummer) und die Anzahl der Elemente pro Gruppe erfasst (Variable gc). Zum Abschluss wird der Datensatz nach Postanschrift, PLZ und Ortsname sortiert.

Input: `REX_Institutes_Dynamic_Filter.dta`

Output: `REX_Institutes_Splitter.dta`
found; level0; score; grp; gc; REXid; REXid_Short; Institution1; Institution2; Institution3; PLZ; PLZ_Short; Ortsname; Ortsname_Short; AGS5; AGS_Short; AGS_tmp; Postanschrift; Strasse; Hausnummer

40 SID zuspielen.do

In diesem Schritt werden die Schreibweisen der Institutsnamen standardisiert und dem Datensatz einige Felder aus der in 10 Institutes_REX.do erzeugten SID-Tabelle `Institutes_REX_SID_suffix.dta` hinzugefügt (Institution_SID; Strasse_SID; Hausnummer_SID; PLZ_SID; AGS5_SID). Die bisherigen Einträge von Institution1 werden dabei in einem zweiten Schritt überschrieben.

Das Ergebnis ist eine Tabelle (`REX_Institutes_SID`), welche die Evaluierung der Zuordnung von langer Version des REX und der in 10 Institutes_REX.do erstellten kurzen, erweiterten Version ermöglicht.

Input: `REX_Institutes_Splitter.dta`
`Institutes_REX_SID_suffix.dta`

Output: `REX_Institutes_SID.dta`
found; level0; score; grp; gc; REXid; REXid_Short; Institution1; Institution2; Institution3; PLZ; PLZ_Short; Ortsname; Ortsname_Short; AGS5; AGS5_SID; AGS_Short; AGS_tmp; Postanschrift; Strasse; Strasse_SID; Hausnummer; Hausnummer_SID; BuLa

4. Tools

In diesem Schritt werden Werkzeuge zur Erstellung zusätzlicher Tabellen verwendet.

Multiple AGS.do

In diesem Schritt werden Institutionen mit Standorten in mehreren Landkreisen (AGS) identifiziert. Typische Fälle sind Universitäten mit mehreren Standorten.

Ausgehend von der SID-Tabelle aus 40 SID zuspielen.do (`REX_Institutes_SID.dta`), wird zuerst die Anzahl der Einträge für jede Gruppierung aus REXid_Short und AGS5 bestimmt (Variable AGS_c).

Anschließend werden alle Einträge gelöscht, die in einer REXid-Gruppe nur das „Haupthaus“ enthalten (level0 = 1; gc = 1 - siehe 20 Dynamic Filter.do). Diese Gruppe von Institutionen hat keine weiteren Standorte. Weiter werden alle Einträge gelöscht, deren AGS (AGS5) der des Haupthauses (AGS5_SID) entspricht.

Nach dem Entfernen von Duplikaten werden ausgehend von der REXid der identifizierten Version (REXid_Short) zusätzliche Daten aus der kurzen Version des REX (`20140321_Forschungseinrichtungen_REX.dta`) nachgeladen. Dies geschieht um z. B. Daten für Einrichtungstyp und Sektion für die untergeordneten Organisationseinheiten in den Kreisen zu übernehmen.

Nun wird eine erweiterte REXid generiert (REXid_ext), die in den ersten 12 Stellen die bisherige ID des „Haupthauses“ enthält (REXid_Short), zusätzlich aber die durch „-“ abgetrennte Angabe der Kreiskennziffer (AGS5). Die resultierende REXid_ext ist somit 18 Stellen lang.

Der resultierende Datensatz ist ein SID-Datensatz (`REX_Multiple_AGS_SID.dta`). Die Schreibweisen der Institutionen stammen aus der SID-Tabelle `REX_Institutes_SID.dta` aus 40 SID zuspielen.do. Diese Daten erhalten die Quellencodierung `source = 50` und stellen neue Einträge für Standorte dar.

Um weitere Schreibweisen für die Institutionen zu erhalten, wird die Tabelle mittels m:m-Verknüpfung (REXid) mit der Tabelle `Institutes_REX_Spellings.dta` aus 10 Institutes_REX.do verknüpft, welche die verschiedenen Schreibweisen für jede REXid enthält. Die so neu hinzugewonnenen Einträge enthalten also alternative Schreibweisen für die Institution und erhalten die Quellencodierung `source = 51`. Diese Daten bilden den Datensatz `REX_Multiple_AGS.dta`. Der Datensatz `REX_Multiple_AGS_Eval.dta` entspricht ebenfalls diesem Stand, enthält jedoch noch weitere Variablen zur Evaluierung der Routine.

***Anmerkung**: Auch hier sei darauf verweisen, dass zur Neuerstellung der hier erzeugten Tabellen das Nachladen von `REX_Multiple_AGS.dta` in Schritt 10 Institutes_REX.do unterbleiben muss.*

Input: `REX_Institutes_SID.dta`
`Institutes_REX_Spellings.dta`
`20140321_Forschungseinrichtungen_REX.dta`

Output: `REX_Multiple_AGS_SID.dta`
AGS_c; REXid; REXid_Short; Institution; AGS5; AGS5_SID; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

Output: `REX_Multiple_AGS_Eval.dta`
AGS_c; REXid_ext; REXid_Short; Institution_Short; Institution; AGS5; AGS5_SID; Strasse; Hausnummer; PLZ; Ortsname; OrtsnamemitZusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

Output: `REX_Multiple_AGS.dta`
REXid; Institution; AGS5; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

Output: `REX_Multiple_AGS.tsv`
REXid; Institution; AGS5; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion; source

Literatur

Ehrenfeld, Wilfried (2015a): RegDemo: Aufbereitung und Zusammenführung der Akteursdaten - Technische Dokumentation der Routinen und Datensätze. IWH Technical Reports 1/2015.

Ehrenfeld, Wilfried (2015b): Research Explorer - Technische Dokumentation der Routinen. IWH Technical Reports 3/2015.

Ehrenfeld, Wilfried (2015c): RLPC: Record Linkage Pre-Cleaning - Technische Dokumentation der Routinen. IWH Technical Reports 2/2015.

Titze, Mirko, Wilfried Ehrenfeld, Matthias Piontek und Gunnar Pippel (2015): „Netzwerke zwischen Hochschulen und Wirtschaft: Ein Mehrebenenansatz“. In: Schrumpfende Regionen - dynamische Hochschulen: Hochschulstrategien im demografischen Wandel. Hrsg. von Michael Fritsch, Peer Pasternack und Mirko Titze. Wiesbaden: Springer Fachmedien. Kap. 11, S. 213–234.

A. Anhang

A.1. Datentypen und Darstellung – Prinzipieller Aufbau

Beispiel: 20140507_Forschungseinrichtungen_REX_AGS.dta

Name	Typ	Formatierung
REXid	str12	%-12s
Postanschrift	strL	%-100s
Institution	strL	%-90s
[Originalschreibweise]	strL	%-80s
[Kommentar]	strL	%-35s
Strasse	str35	%-50s
Hausnummer	str17	%10s
PLZ	str5	%-5s
Ortsname	str24	%-24s
OrtsnamemitZusatz	str29	%-29s
AGS5	str5	%-5s
Bundesland	str22	%-22s
Internetadresse	strL	%-35s
Fachgebiet	strL	%-35s
Einrichtungstyp	byte	%-35.0g
Sektion	byte	%-35.0g
[source]		%6.0g
[len]		%6.0g

A.2. Codierung Einrichtungstypen

Die Variable Einrichtungstyp enthält folgende Angaben zum Typ der Einrichtung:

Code	Bedeutung
1	Wirtschaft
2	Hochschulen
3	Außeruniversitäre Forschung
4	Akademien der Wissenschaft
5	Ressortforschung von Bund und Ländern
6	Sonstige Forschungseinrichtungen
7	Sonstige
8	Natürliche Personen bzw. Privatpersonen [aus DPMA]

A.3. Codierung Sektionen

Die Variable Sektion enthält folgende Angaben zum Typ der Einrichtung:

Code	Bedeutung
1	Wirtschaft
21	Fachhochschulen
22	Musik- und Kunsthochschulen
23	Universitäten
31	Fraunhofer-Gesellschaft
32	Helmholtz-Gemeinschaft
33	Leibniz-Gemeinschaft
34	Max-Planck-Gesellschaft
4	Akademien der Wissenschaft
51	Bundesforschungseinrichtungen
52	Landesforschungseinrichtungen
6	Sonstige Forschungseinrichtungen
61	Bibliotheken und Archive (ohne Hochschulbibliotheken)
62	Deutsche Fördereinrichtungen
63	Krankenhäuser, Kliniken und Therapiezentren (ohne Universitätskliniken)
7	Sonstige

A.4. Codierung Datenquelle

Die Variable source enthält folgende Angaben zur Quelle der Daten:

Code	SID	Bedeutung
0	*	Einträge, die umbenannt werden sollten [REX_ADD_Rename]
1	*	Die komplette kurze Version des Research Explorers [20140321_Forschungseinrichtungen_REX]
2		Zusätzliche Einträge, bei denen die Schreibweise für Institution in der langen Version von der kurzen Version abweicht [REX_ADD_Same_ID]
3	*	Zusätzliche Einträge für hinzugefügte Einträge aus der langen Version sowie Aliase der kurzen und langen Version [REX_ADD_Long_Alias]
31		Alias für Eintrag aus der kurzen Version
32	*	Übernommener Eintrag aus der langen Version
33		Alias für Eintrag aus der langen Version
34		<nicht verwendetes Spezialfeld zum testen>
4	*	Zusätzliche Einträge für manuell erfasste Standorte [REX_ADD_Standorte]
40	*	Neuer Eintrag für Standort
41		Alias eines Standortes
5	*	Zusätzliche Einträge mit multipler AGS pro REXid [REX_Multiple_AGS]
50	*	Neuer Eintrag für Standort
51		Alias eines Standortes

A.5. Code Statistics

Stand: August 2015

Modul	Anzahl Zeilen
2 Aufbereitung	
01 Convert 20140321_Forschungseinrichtungen_REX.do	115
02 Convert 20140507_Forschungseinrichtungen_REX.do	153
03 Prepare REX_ADD_Long_Alias.do	435
04 Prepare REX_ADD_Same_ID.do	182
05 Prepare REX_ADD_Standorte.do	252
06 Prepare REX_ADD_Rename.do	139
07 Prepare REX_ADD_Delete.do	85
3 Zusammenfügung	
10 Institutes_REX.do	455
20 Dynamic Filter.do	273
20 Program_Filter_Institutions.do	325
30 Splitter.do	277
40 SID zuspielen.do	112
Tools	
Multiple AGS.do	270
#1 Prepare datasets.do	20
Anzahl Module	14
Codezeilen Gesamt	**2958**

Leibniz-Institut für Wirtschaftsforschung Halle – IWH

HAUSANSCHRIFT: Kleine Märkerstraße 8, D-06108 Halle (Saale)

POSTANSCHRIFT: Postfach 11 03 61, D-06017 Halle (Saale)

TELEFON: +49 345 7753 60 TELEFAX +49 345 7753 820

INTERNET: www.iwh-halle.de
I S S N : 2 3 6 5 - 9 0 7 6