Halle Institute
for Economic Research
Member of the Leibniz Association

IWH TECHNICAL REPORTS

Research Explorer

Technical Documentation of Routines

Wilfried Ehrenfeld

03|2015

Author:
Dr Wilfried Ehrenfeld

Contact:
Dr Cornelia Lang
Coordinator of the IWH Data Centre
Phone: + 49 345 77 53 802
Fax: + 49 345 77 53 820
E-mail: cornelia.lang@iwh-halle.de

Editor: HALLE INSTITUTE FOR ECONOMIC RESEARCH (IWH) – MEMBER OF THE LEIBNIZ ASSOCIATION
Executive Board: Professor Reint E. Gropp, PhD
Professor Dr Oliver Holtemöller
Dr Tankred Schuhmann

Adress: Kleine Maerkerstrasse 8, D-06108 Halle (Saale), Germany
Postal Address: P.O. Box 11 03 61, D-06017 Halle (Saale), Germany
Phone: +49 345 7753 60
Fax: +49 345 7753 820
Internet: www.iwh-halle.de

Citation:
Ehrenfeld, Wilfried: Research Explorer – Technical Documentation of Routines. IWH Technical Reports 03/2015. Halle (Saale) 2015.

ISSN 2365-9076

Research Explorer

Technical Documentation of Routines

Abstract

The Research Explorer is a directory of institutes at German universities and non-university research facilities, which is provided by the German Research Foundation (DFG) and the German Academic Exchange Service (DAAD). It is available for download for free and currently comprises ca. 23,000 entries. Two versions of this directory, which differ with regard to depth and coverage, are available to us. One version comprises 1712 entries, while the other holds over 22,000.

Primary goal of the presented routines is the comparison and complement of these two versions of the Research Explorer. The necessity for this arises from the differences in depth and coverage of the different versions. Furthermore, both versions are incomplete at some points, which means that individual institutions had to be added manually. The resulting data sets of the modified Research Explorer can subsequently be used for the mapping of cooperation relations.

Contents

List of Figures

List of Tables

1. Starting basis and outline of the procedure

The Research Explorer (short: REX) is a directory of institutes at German universities and non-university research facilities. It is a joint product of the German Research Foundation (DFG) and the German Academic Exchange Service (DAAD) in cooperation with the German Rectors' Conference (HRK). This directory is provided free of charge as online data base[1] and comprises almost 23,000 entries. The information is organized according to geographical, technical and structural criteria and, in addition to the names of the institutions, include details on the place (variables: Strasse (street); Hausnummer (street number); PLZ (zip code); Ortsname (city); Bundesland (state) and type (variables: Fachgebiet (field); Einrichtungstyp (facility type); Sektion (branch)).

Two versions of this directory, which differ with regard to depth and coverage, are available to us. We can distinguish between the "short" or "small" version, and the "long" or "large" version of the REX. The "short" version comprises 1,712 entries (as of 21.03.2014). The "long" version contains 22,331 entries (as of 07.05.2014). In the latter, the subdivision of universities is depicted down to the level of professorial chairs, whereas in the former only the university is listed as a superordinate unit. Furthermore, the long version also includes university hospitals, which the short version is lacking entirely.

For both versions of the Research Explorer the identification numbers for professorial chairs, places of research facilities and similar structures was not assigned hierarchically, but independently. For research purposes, however, a hierarchical system of IDs would have been desirable. Additionally, notations of institutions differ between the short and the long version in several cases. Furthermore, both versions are incomplete at some points, which means that individual institutions had to be added manually. This mainly applies to sites of the Fraunhofer Society and other non-university research facilities.

Therefore, the primary goal of the presented routines is the comparison and complement of these two versions of the Research Explorer. The necessity for this arises from the differences in depth and coverage of the different versions. This does not mean that entries of the databases should simply be compared. Rather, the short version should be complemented with a shortened hierarchical system of locations from the long version. The county code was chosen as aggregation property.

This is why a filter has been developed for the realization. On the basis of an expanded short version of the REX, this filter checks all entries of the long version for an affiliation of the institution with an institution from the expanded short version. The resulting data sets of the modified Research Explorer can subsequently be used for mapping cooperative relations (see Titze et al. 2015 and Ehrenfeld 2015a) by means of record linkage methods (see Ehrenfeld 2015c), for example in conjunction with data from the company database Amadeus. Figure 1 depicts the structural sequence of the process. The following describes the individual stages of the procedure more in-depth.

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

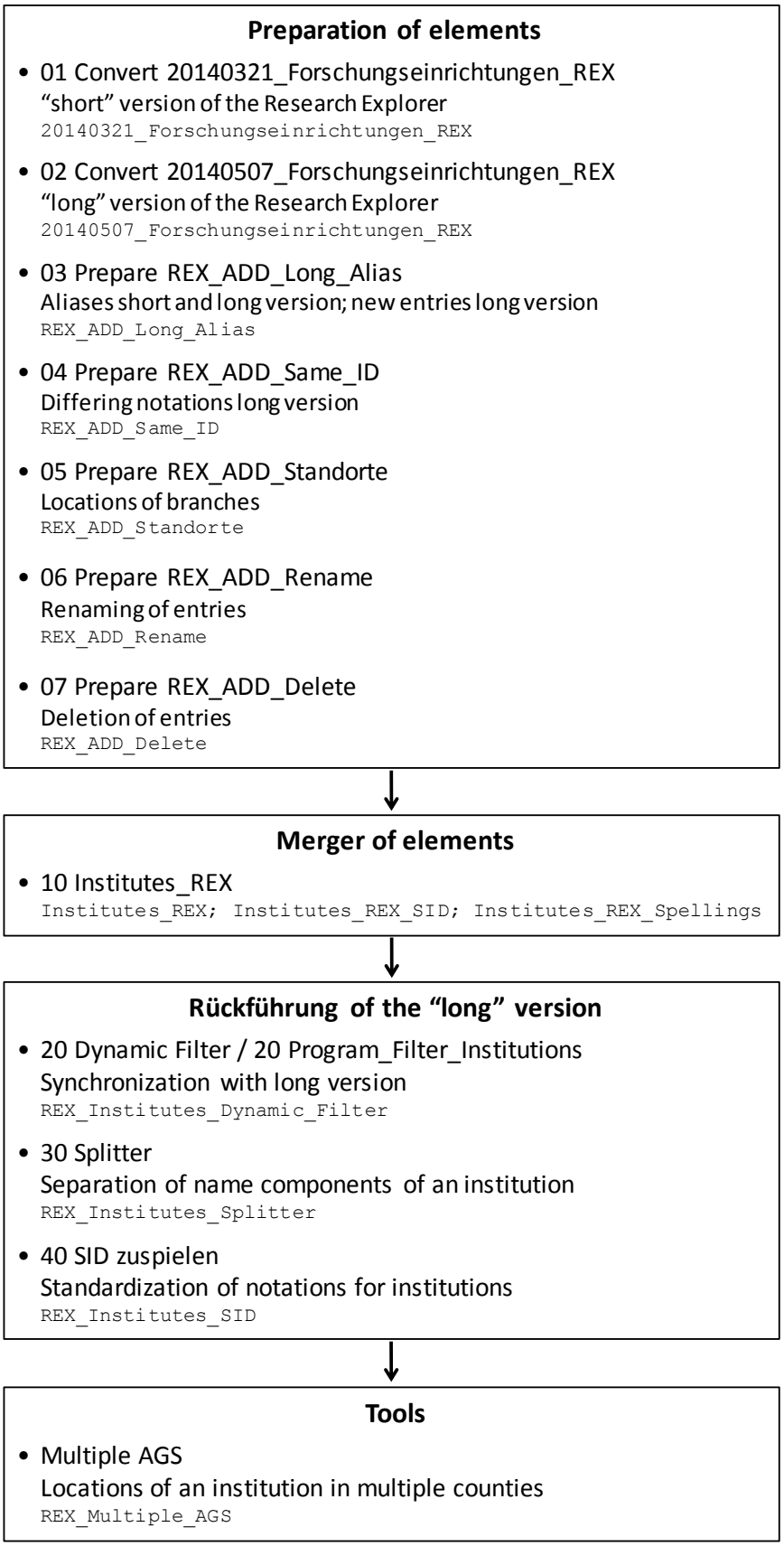

Figure 1: Flowchart and resulting data sets of the applied procedure.

2. Processing the individual components

In this step the individual components of the expanded REX data set are read and treated.

01 Convert 20140321_Forschungseinrichtungen_REX.do

The data from the small version of the Research Explorer is read from the xlsx file. The existing Id field is converted into a format with constant length (REXid). Here, the ID comprises 12 characters, the first three of which represent the source (REX). The following 9 characters are numerically identical to the number from the original Id field.

The fields are adjusted for bothering control characters (`CR`; `LF`; `TAB`; `QUOTES`) and superfluous (i.e. at the beginning or the end) and double spaces. Einrichtungstyp and Sektion are converted from strings into numerical values and labeled (see appendix A.2 and A.3). The individual variables are adjusted, formatted (see appendix A.1) and sorted (sorting order: 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

This routine performs the same tasks as 01 Convert 20140321_Forschungseinrichtungen_REX.do does for the large Research Explorer data set. Additionally, another version is created. For this new version the county code (AGS5) is derived from PLZ and Ortsname through an external routine (`PLZ_AGS_Prog`).

In addition to Postanschrift, the long version contains the field Institution. In contrast to the field with the same name in the short version, this field contains details on subordinate organizational elements like professorial chairs.

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

The table `REX_ADD_Long_Alias.xlsx` captures three types of adjuncts:

- Aliases of institutions from the short version [`Status = 1`]
- New entries, taken over from the long version [`Status = 2`]
- Aliases of institutions from the long version [`Status = 3`]

The routine reads in the Excel file and processes it (generating REXid; setting variable types; adjusting for spaces). This also includes the capitalization of institution names (Originalschreibweise; Postanschrift; Institution) and Ortsname (external routine `WEupper`). In the course of this, the name of the Institution is renamed to Institution_add. Subsequently, REXid is used to add missing data from the short version or the long version.

Firstly, the data set is complemented with details from the short version on the basis of REXid. The variable `_merge` is used to determine that the added data are from the short version. Then, the variable Status states the source of the additions. For the records identified and complemented this way the Status is set to: `Status = 1`. The records not yet allocated get `Status = 0`.

By comparing the field Originalschreibweise with Institution from the short version, it is subsequently established whether the entry is a valid alias for an already existing entry. The variable test states the result of the comparison of the aliases with the Institution from the short version. It can take on three states:

- No existing original notation [`test = 0`]
- An original notation exists but differs [`test = 1`].
  It should therefore be corrected
- The original notation is okay [`test = 2`]

Following this, the data set is linked with the long version. The identification of new entries taken over from the long version is done on the basis of the institution name (Institution_add) with the field Postanschrift, subject to the condition that Originalschreibweise is empty. For the records identified this way Status is set to `Status = 2`. Here, again, the variable test states the result regarding Originalschreibweise. For the newly added records from the long version Originalschreibweise has to be empty (`test = 2`).

Finally, the aliases from the long version are assigned `Status = 3`. As for the aliases for the short version, the variable test states the result of the comparison of notations. For this, the institution (Institution_add) must not be identical to the field Postanschrift.

Finally, there is an assessment to make sure that for each alias of the long version there is a corresponding entry in the long version. root and root2 are used for this. Variable root is equal to 1 if the entry is an alias from the long version (`Status = 3`), otherwise it is 0. Variable root2 is defined as the minimum of root for each (group of) REXid. This means that each group of REXid must also contain the newly added entries from the long version (`Status = 2`). For these, however, `root = 0` applies. If the minimum of root2 is 0 the alias in the REXid-group has a corresponding root-entry from the long version. If `root2 = 1`, such an entry is missing. This has to be corrected.

The result of this routine are three data sets:

- one data set to evaluate the results of this routine. This one also contains all aliases and adopted entries from the long version [`REX_ADD_Long_Alias_Eval.dta`].
- one data set with all aliases and the new entries adopted from the long version. In this data set the variable Status is renamed (source) and modified in order to capture the source of the complementing data set (`Status = 1` → `source = 31` - see appendix A.4) [`REX_ADD_Long_Alias.dta`].
- one SID data set[2]which only contains the newly added data from the long version (`Status = 2`) [`REX_ADD_Long_Alias_SID.dta`].

[2] SID means: **S**ingle **ID**. In this data set there exists exactly one entry for each 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

Given the same REXid, this routine recognizes different notations for Institution in the long (suffix: _long) and the short version (suffix: _short) of the REX. Names of institutions and places are capitalized (external routine `WEupper`). Identification is done by means of a scoring system. For this, the Institutionname (`score = 4`), the postal code (PLZ; `score = 2`) and the place name (Ortsname; `score = 1`) are taken into account. If at least one of these properties is deviating, the entry is added to the supplementary table (here, however, an examination shows that `score` is either 0 or 7). The data is isolated, treated, sorted and given a source indicator (`source = 2`; see appendix A.4).

Compared to `REX_ADD_Same_ID`, the file `REX_ADD_Same_ID_Eval` contains additional variables, which are used for evaluation of the routines.

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

The file `REX_ADD_Standorte.xlsx` contains locations of institutions already present in the REX, as well as their different notations. For this, a hierarchical ID system was created

on the basis of the (pre-existing) IDs of the "main locations". Each (usually subordinate) branch is allocated the ID of the "main location", which is extended with a suffix. A hyphen ("-") is added to separate the three-digit suffix, which represents the sequential branch number, from the main ID. Consequently, these REXid fields comprise 16 characters.

After the data has been read from the Excel table, Institution is converted into upper case (external routine `WEupper`). The variables are processed, duplicates are deleted and institution types are allocated (external routines `WEreplace_Einrichtungstyp`; `RDlabel_Einrichtungstypen`; `RDlabel_Sektionen` - see appendices A.2 and A.3). Required data is added from the short version. Strasse and Ortsname are also capitalized.

Subsequently, the variable Status, which states whether an entry is an alias for a location, is created. Should an entry be an alias of an existing entry, it holds: `Status = 1`, else 0. The check is done on the basis of the content of Originalschreibweise. If this field is not empty, the entry is an alias.

Finally, the data is labeled with a source indicator that is derived from Status (`source = 40` bzw. `source = 41`; see appendix A.4). Table `REX_ADD_Standorte_Eval.dta` contains additional information for the evaluation of the routine. All aliases are deleted for the SID table `REX_ADD_Standorte_SID.dta`.

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

`REX_ADD_Rename.xlsx` contains notations from the short REX version that are to be changed. This is necessary if the notation given there does not match the actual state. Nevertheless, the "wrong" notations should be maintained in order to be able to recognizes these entries when comparing with the long version.

The data from the Excel table is read in. Originalschreibweise and Institution are capitalized. Institution is renamed to Institution_ren in order to be able to add data from the short version. The data are complemented and the references of the aliases to the respective institutions are checked for correctness.

If Originalschreibweise matches the field Institution from the short version(!), the alias is valid and it holds: `test = 1`, else 0. Furthermore, it is checked whether the Institution field from the short version matches the institution field (Institution_ren) from the rename-table (Variable test2). Should this be the case (test2 = 1), adding the replacement would be pointless since nothing has to be renamed.

The Eval-table `REX_ADD_Rename_Eval.dta` contains further information for the evaluation of the 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

Table `REX_ADD_Delete.xlsx` provides a "droplist". All entries in it are to be deleted from the data set later on. The necessity for such an approach is based on two facts. Firstly, inclusion of locations of research institutes in the Research Explorer is incomplete (see also sections 05 Prepare REX_ADD_Standorte.do and Multiple AGS.do). Secondly, entries have been included multiple times with slightly varying notations (or German/English).

The data are read from the Excel table `REX_ADD_Delete.xlsx`. The 9-digit REXid is created from the ID field. Institution is capitalized. The data are adjusted and formatted.

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

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

3. Merger of the data sets and comparison with the long version

In this step the prepared components of the data set are merged and finalized.

10 Institutes_REX.do

Here, the data sets created and extracted so far are merged. In the course of this three groups of data sets, which are needed later on, are created:

- `Institutes_REX.tsv` is the resulting search filter, which is subsequently used to search through the long version of the REX (see 20 Dynamic Filter.do). The `dta` version is the same as the `tsv` version with regard to the contents.
- **`Institutes_REX_Eval.dta`** can be seen as the first main result of the procedure. It contains all REXids with all notations and therefore represents the expanded (short) version of the REX as identification table.
- **`Institutes_REX_SID.dta`** is the procedure's second main result. Just like `Institutes_REX_Eval.dta`, it contains all REXid entries, but with the difference that for each REXid entry only one notation is registered. It is an SID list with unique REXid numbers and will later on be used for the standardization of identified entries (see 40 SID zuspielen.do).
- `Institutes_REX_SID_suffix.dta` is the same list, but the suffix `_SID` is added to the fields.
- `Institutes_REX_Spellings.dta` contains the different notations for Institution for a REXid. It's coverage is the same as that of the table `Institutes_REX.dta`, but it only contains the REXid and the Institution. It is used in Multiple AGS.do to create alternative notations for locations in multiple counties.

The data set is merged by using `append` with the read-in and prepared data sets from section 2. Furthermore, the data sets of institutions with multiple locations in different counties created in Multiple AGS.do are added. These data are given the source indicator `source = 50` or `source = 51` (see appendix A.4).

Note: *If the table `REX_Multiple_AGS.dta` created in Multiple AGS.do is to be generated anew, the inclusion of this data set (`append`) has to be deactivated at this point and the rest of the routine has to be run again. Multiple AGS.do will subsequently create a new `REX_Multiple_AGS.dta`.*

The data from the short version of the REX (`20140321_Forschungseinrichtungen_REX.dta`) are given the source indicator `source = 1` (see appendix A.4). The data from the rename-file (`REX_ADD_Rename.dta`) are given the source indicator `source = 0`. The coding of the sources in the variable source corresponds to appendix A.4.

The delete-table (`REX_ADD_Delete.dta`) is added (`merge m:1 REXid`) and the data sets identified this way are deleted. The data are adjusted, Institution, Ortsname and Strasse are capitalized. The data set is sorted by the variables Institution, REXid, PLZ and Ortsname, and duplicates of these variable are identified (dup), marked for deletion (drp) and eventually deleted.

Finally, the county code (AGS5) is derived (external routine PLZ_AGS_Prog) from the postal code (PLZ) and the place name (Ortsname), the variables are sorted and the length of the string field Institution is determined (len). This list is sorted by this length in descending order for the filter table Institutes_REX.tsv. The consequence of this is that special cases (e.g. Universität ... Klinikum) are found before the more general cases (Universität ...) when filtering later on.

Records from sources with the source-codes 2 (alias long vs. short version), 31 (alias short version), 33 (alias long version), 41 (alias location) and 51 (alias multiple AGS) are deleted for the SID table (Institutes_REX_SID.dta; see appendix A.4). Code 32 is changed to code 3 in the course of this. Duplicates regarding REXid are deleted.

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

The long version of the Research Explorer (20140507_Forschungseinrichtungen_REX.dta) is read in. The data set is treated, Postanschrift, Institution, Strasse and Ortsname

are capitalized. Several flags and temporary variables are created in preparation for the filtering (table 1; further explanation in 20 Program_Filter_Institutions.do).

Table 1: Flags and temporary variables in 20 Dynamic Filter.do

Name	Type	Explanation
found	byte	Flag: Entry found
level0	byte	Flag: Perfect match: `Inst_Short` = `Inst_Long`
subst	byte	Flag: Replace data set
score	byte	Flag: Data set match Short Long
score_tmp	byte	Temp: Data set match Short Long
REXid_Short	str12	REXid of short version
REXid_tmp	str12	Temp: REXid of short version
REXid_hist	str12	Temp: Replacement history of REXid
Inst_Long	strL	Duplicate of Postanschrift
Inst_Short	strL	Institution of short version
Inst_tmp	strL	Temp: Institution of short version
PLZ_Short	str5	PLZ of short version
PLZ_temp	str5	Temp: PLZ of short version
Ortsname_Short	str24	Ortsname of short version
Ortsname_tmp	str24	Temp: Ortsname of short version
AGS_Short	str24	AGS5 of short version
AGS_tmp	str24	Temp: AGS5 of short version

The variables are formatted and brought into the right order. The institution field from the long version is renamed to Institution3. The filter routine 20 Program_Filter_Institutions.do is applied and the distribution of score is displayed. This filter is used to allocate matching entries from the extended short version of the REX (`Institutes_REX`) created in 10 Institutes_REX.do to entries from the long version. After the filtering the allocated data can be found in the fields with the suffix _Short: REXid_Short; Inst_Short; PLZ_Short; Ortsname_Short und AGS_Short).

The variable level0 is given the value 1 if `Inst_Short` = `Inst_Long` holds, else it is 0. It indicates that the record refers to the superordinate institution itself, meaning that Inst_Long does not include any further information on subordinate organizational elements. Finally, the temporary variables (suffix _tmp) are deleted.

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

This program contains the dynamic filter for comparing the long version of the REX with the expanded short version. The filter file `Institutes_REX.tsv` created in 10 Institutes_REX.do is opened and the first line is read. The fields of the tsv file separated by the TAB-symbol are split off one after another and transferred into local variables: ID; Inst; PLZ; Ort and AGS

Candidates for the allocation of institutions are identified by comparing the variable Inst_Long (which is a duplicate of Postanschrift) with the local variable Inst (from the tsv file). Such a candidate could possibly be found if the local variable Inst stands at the beginning of Inst_Long. In this case the counter for matches found (Flag found) is increased by one and the fields from the tsv file (ID; Inst; PLZ; Ort and AGS) are adopted by the temporary variables REXid_tmp; Inst_tmp; PLZ_tmp; Ortsname_tmp and AGS_tmp. Subsequently, a scoring system (see table 2) is used to determine the goodness of fit of different variables, which is then allocated to the temporary variable score_tmp.

Table 2: Scores in 20 Program_Filter_Institutions.do

score	Explanation
32	Perfect match between institution names Inst_Long und Inst_tmp.
16	Inst_tmp stands at the beginning of Inst_Long and is longer than the previous entry of Inst_Short. This check is done in order to give extended entries like Universität ... Klinikum a higher weight than their superordinate and therefore shorter institutions (Universität ...)
8	Inst_tmp stands at the beginning of Inst_Long
4	Match between PLZ and PLZ_tmp
2	Match between Ortsname and Ortsname_tmp
1	Match between AGS5 and AGS_tmp

In case the record has a higher score in the _tmp-variables (score_tmp) than the best candidate so far has in the _Short-variables (score), the flag subst is given the value 1 (else 0). The _Short-record is subsequently overwritten with the _tmp-record and the previous REXid_Short is added to the variable REXid_hist.

Then the next line of Institutes_REX.tsv is read-in. This is done until there is no more data left in the tsv file. Finally, the tsv file is closed.

Input: `Institutes_REX.tsv`

30 Splitter.do

Following the filtering, this routine separates the additional information from the institution name of the long version (Inst_Long). The aim of this splitter is a hierarchical structure of institution levels according to Institution1, Institution2 and Institution3.

In the long version of the REX there is, in addition to Postanschrift (which can also be found in Inst_Long as working copy), the field Institution. In contrast to the field with the same name in the short version, this field contains details on structurally subordinate organizational elements, e.g. chairs. This field was renamed to Institution3 in 20 Dynamic Filter.do. It contains the lowest subordinate hierarchy level.

Firstly, the field Institution3 is emptied if it's content matches Inst_Long. Subsequently, the superordinate institution identified in Inst_Short is removed from Inst_Long if this term is standing at the beginning. The rest of Inst_Long then holds the remaining middle and last hierarchy level. Consequently, Inst_Short is renamed to Institution1 and Inst_Long is renamed to Institution2.

Furthermore, Institution3 is emptied if it's content is identical to Institution2. In case the complete content of Institution3 is located at the end of Institution2, this part is removed from Institution2. Finally, Institution3 is emptied if it's content is already included in Institution2.

The entries are grouped (variable grp contains the group number) by the REXid of the short version (REXid_Short) and the number of elements for each group are captured (variable gc). In the end the data set is sorted by Postanschrift, PLZ and Ortsname.

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 this step the notations for institution names are standardized and several fields from the SID table `Institutes_REX_SID_suffix.dta` created in 10 Institutes_REX.do are added to the data set (Institution_SID; Strasse_SID; Hausnummer_SID; PLZ_SID; AGS5_SID). Previous entries of Institution1 are overwritten in a second step.

The result is a table (`REX_Institutes_SID`), which allows the evaluation of the allocation of the long version of the REX and the short, expanded version created in 10 Institutes_REX.do.

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 this step tools for the creation of additional tables are used.

Multiple AGS.do

In this step, institutions with locations in multiple counties (AGS) are identified. Typical cases are universities with several locations.

Firstly, the number of entries for each group from REXid_Short and AGS5 (Variable AGS_c) is determined on the basis of the SID table from 40 SID zuspielen.do (`REX_Institutes_SID.dta`).

Subsequently, all entries only containing the "main location" of a REXid group are deleted (level0 = 1; gc = 1 - see 20 Dynamic Filter.do). This group of institutions does not have further locations. Furthermore, all entries whose district code (AGS5) matches the one of the main location (AGS5_SID) are deleted.

Following the removal of duplicates, further data from the short version of the REX (`20140321_Forschungseinrichtungen_REX.dta`) are reloaded based on the REXid of the identified version (REXid_Short). This is done in order to, for example, adopt data on the facility type (Einrichtungstyp) and branch (Sektion) for the subordinate organizational units in the counties.

Now an extended REXid is generated (REXid_ext). It contains the previous ID of the "main location" (REXid_Short) in the first 12 characters and, in addition to that, the county code (AGS5) separated by a "-". In consequence, the resulting REXid_ext comprises 18 characters.

The resulting data set is a SID data set (`REX_Multiple_AGS_SID.dta`). The notations for institutions are from the SID table `REX_Institutes_SID.dta` from 40 SID zuspielen.do. These data are given the source code `source = 50` and represent new entries for locations.

In order to get further notations for institutions, the table is linked with the table `Institutes_REX_Spellings.dta` from 10 Institutes_REX.do, which contains different notations for each REXid, by means of an m:m assignment (REXid). The newly acquired entries therefore contain alternative notations for the institution and are given the source code `source = 51`. These data form the data set `REX_Multiple_AGS.dta`. The data set `REX_Multiple_AGS_Eval.dta` also represent this state of the data, but contains further variables for the evaluation of the routine.

***Note**: It should again be pointed out that, if the tables created here are to be regenerated, `REX_Multiple_AGS.dta` must not be loaded in step 10 Institutes_REX.do.*

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

References

Ehrenfeld, Wilfried (2015a): RegDemo: Preparation and Merger of Actor Data - Technical Documentation of Routines and Datasets. IWH Technical Reports 1/2015.

Ehrenfeld, Wilfried (2015b): Research Explorer - Technical Documentation of Routines. IWH Technical Reports 3/2015.

Ehrenfeld, Wilfried (2015c): RLPC: Record Linkage Pre-Cleaning - Technical Documentation of Routines. IWH Technical Reports 2/2015.

Titze, Mirko, Wilfried Ehrenfeld, Matthias Piontek, and Gunnar Pippel (2015): "Netzwerke zwischen Hochschulen und Wirtschaft: Ein Mehrebenenansatz." In: Schrumpfende Regionen - dynamische Hochschulen: Hochschulstrategien im demografischen Wandel. Ed. by Michael Fritsch, Peer Pasternack, and Mirko Titze. Wiesbaden: Springer Fachmedien. Chap. 11, pp. 213–234.

A. Appendix

A.1. Data types and presentation – general structure

Example: 20140507_Forschungseinrichtungen_REX_AGS.dta

Name	Type	Format
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. Coding of institution types

The variable Einrichtungstyp contains the following details on the type of an institution:

Code	Explanation
1	Business
2	Universities
3	Non-university research
4	Academies of Science
5	Departmental research of federal government and federal states
6	Other research facilities
7	Other
8	Natural or private persons [from DPMA]

A.3. Coding sections

The variable Sektion contains the following details on the type of an institution:

Code	Explanation
1	Business
21	Universities of Applied Sciences
22	Music and Art Academies
23	Universities
31	Fraunhofer Society
32	Helmholtz Society
33	Leibniz Society
34	Max-Planck Society
4	Academies of Sciences
51	Federal research institutes
52	County research institutes ("Landesforschungseinrichtungen")
6	Other research facilities
61	Libraries and archives (excluding university libraries)
62	German funding organizations
63	Hospitals, clinics and therapy centers (excluding university hospitals)
7	Other

A.4. Coding data source

The variable source contains the following details on the source of data:

Code	SID	Explanation
0	*	Entries to be renamed [REX_ADD_Rename]
1	*	The complete short version of the Research Explorer [20140321_Forschungseinrichtungen_REX]
2		Additional entries for which the notation of Institution in the long version differs from the one in the short version [REX_ADD_Same_ID]
3	*	Additional entries for added entries from the long version as well as aliases from the short and long version [REX_ADD_Long_Alias]
31		Alias for entry from short version
32	*	Adopted entry from long version
33		Alias for entry from long version
34		<unused special field for testing purposes>
4	*	Additional entries for manually captured locations [REX_ADD_Standorte]
40	*	New entry for location
41		Alias of a location
5	*	Additional entries with multiple AGS per REXid [REX_Multiple_AGS]
50	*	New entry for location
51		Alias of a location

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**

Halle Institute for Economic Research (IWH) – Member of the Leibniz Association

ADDRESS: Kleine Maerkerstrasse 8, D-06108 Halle (Saale), Germany

POSTAL ADDRESS: P.O. Box 11 03 61, D-06017 Halle (Saale), Germany

PHONE: +49 345 7753 60 FAX +49 345 7753 820

INTERNET: www.iwh-halle.de
ISSN: 2365-9076