Ehrenfeld, Wilfried

Research Report

RegDemo: Preparation and Merger of Actor Data – Technical Documentation of Routines and Datasets

IWH Technical Reports, No. 01/2015e

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

Suggested Citation: Ehrenfeld, Wilfried (2015) : RegDemo: Preparation and Merger of Actor Data – Technical Documentation of Routines and Datasets, IWH Technical Reports, No. 01/2015e, Leibniz-Institut für Wirtschaftsforschung Halle (IWH), Halle (Saale)

This Version is available at:
https://hdl.handle.net/10419/144717

Leibniz-Institut für
Wirtschaftsforschung
Halle

IWH TECHNICAL REPORTS

RegDemo: Preparation and Merger of Actor Data

Technical Documentation of Routines and Datasets

Wilfried Ehrenfeld

01|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: RegDemo: Preparation and Merger of Actor Data – Technical Documentation of Routines and Datasets. IWH Technical Reports 01/2015. Halle (Saale) 2015.

ISSN 2365-9076

RegDemo: Preparation and Merger of Actor Data

Technical Documentation of Routines and Datasets

Abstract

Primary objective of the presented routines is the mapping of cooperative relations of companies, universities, non-university research facilities and other institutions on three levels of innovation activity (joint projects; publications, patents). This includes a) the standardization and merging of the three innovation-related databases (funding catalog ("Förderkatalog"); Web of Knowledge; DPMA patents) and b) linking this combined data pool with the data from the institution data sets Amadeus and Research Explorer by means of record linkage procedures. For this project the merger comprises the six regions considered in the RegDemo project. Spatial planning regions ("Raumordnungsregionen") are used for delimitation: 501 - Aachen; 513 - Siegen; 602 - Nordhessen (= "Kassel"); 1302 – Mittleres Mecklenburg/Rostock (= "Rostock"); 1401 - Oberes Elbtal/Osterzgebirge (= "Dresden"); 1504 - Magdeburg.

Contents

List of Figures

1. Problem and procedure description

A realistic depiction of cooperative relations between actors is not possible with a limitation to just one kind of relations, like patents, publications or shared external funds. In order to mitigate this problem a multi-level approach was developed as part of the project "RegDemo". This approach explicitly includes the actors' cooperative relations on the individual levels and combines them (see Titze et al. 2015). To achieve this, the different innovation or cooperation related data sets have to be systematically harmonized and combined.

The following sections describe the used data sets and the chosen procedure. The analyzed database comprises two institution data bases (Amadeus and Research Explorer) and three cooperation related data sets (funding catalog ("Förderkatalog"); Web of Science; German patents - see figure 1). The utilized procedure is based on record linkage methods (see, for example, Christen 2012 and Magerman, Van Looy, and Song 2006).

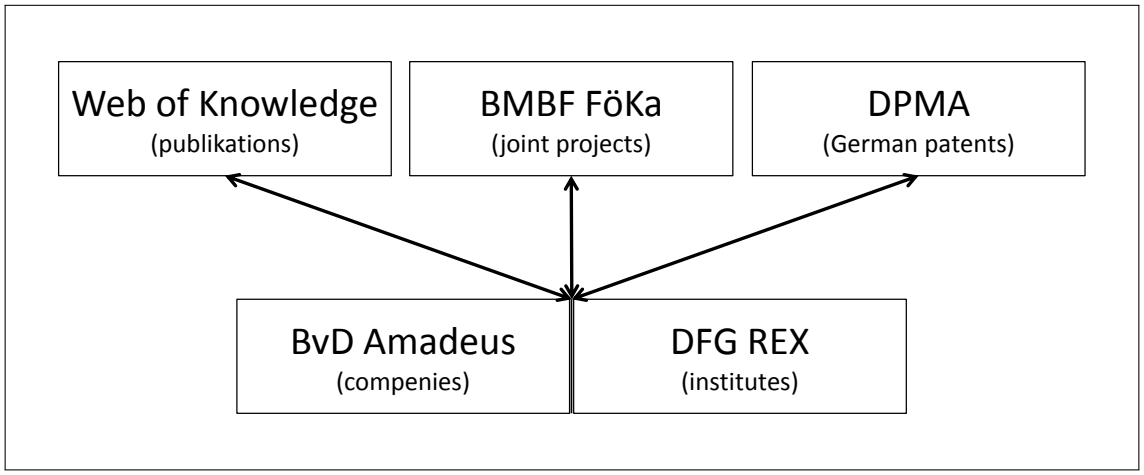

Figure 1: Overview of the data sets linked for the analysis.

1.1. Description of the data sets

Amadeus

The Amadeus data base of the commercial provider Bureau van Dijk comprises information on companies in Europe with currently 14 million entries. Its contents include balance sheet and profit and loss account parameters, other financial parameters and information on the corporate structure of the included enterprises. For Germany the data base includes about 1.5 million companies. Every company in this data set holds a unique identification number (BvD-ID), which will later be used to identify and link actors. A historical Amadeus data set was constructed in order to be able to associate different spellings used in the past for existing companies and to identify companies which no longer exist. It comprises ca. 2.9 million entries for the period 2003-2014. The analysis is restricted to the data sets for the six regions under consideration to speed up the processing later on. The remaining data

set comprises 113,574 entries. This procedure greatly accelerates proceedings without any observed adverse effects on the quality of the allocation.

Research Explorer

The Research Explorer (short: REX) is provided free of charge as online data base[1]. This directory comprises ca. 23,000 entries on institutes at German universities and non-university research facilities. For universities the subdivision is depicted down to the level of professorial chairs. The information is organized according to geographical, technical and structural criteria. This data base complements the above mentioned enterprise data base with regard to cooperation actors, since it usually does not include universities and research facilities.

Funding catalog

The funding catalog ("Förderkatalog") of the Federal Ministry of Education and Research (BMBF) lists data on almost 99,000 current and completed research projects funded by German federal ministries. The funded projects include close to 11,000 joint projects for Germany. For the six regions under consideration 2,416 data sets for 914 joint projects are registered. The recorded information includes the entity funded, time frame, title and thematic reference of the project. Joint projects hold unique numbers ("Verbundnummern"). Every single funding process of a project is recorded with an individual process number. For this analysis we consider all funding projects in the case regions from 1991 to 2010, which have a reference to research and development.

DPMA data base

The data base of the German Patent and Trade Mark Office (DPMA) contains framework data for German patents. The details include an unique patent number and the patent title as well as names and regional information for both the inventor and the applicant. This analysis uses 1,371 co-patents which have been registered between 1994 and 2010, and which fulfill the following criteria: a) At least one applicant or inventor is from one of the relevant case regions; b) at least one cooperative relationship exists within the respective case region under consideration.

Web of Science

Nowadays, the online citation data base Web of Science (formerly ISI Web of Knowledge) is run by the commercial provider Thomson Reuters. The following packages were available for analysis: Science Citation Index Expanded (SCI-EXPANDED); Arts & Humanities Citation Index (A&HCI); Social Sciences Citation Index (SSCI).

[1] see http://research-explorer.dfg.de/research_explorer.de.html.

The evaluated data on co-publications in the case regions cover the period from 2000 to 2012 and comprise 12,502 publications.

1.2. Approach

The objective of this procedure is a) the merger of the three cooperation data sets (funding catalog ("Förderkatalog"); Web of Knowledge; DPMA patents) and b) linking this combined data pool with the data from the institution data sets Amadeus and Research Explorer so as to be able to clearly identify the actors. This method is chosen in order to create a standard for the normalization of notations of the actors and to be able to add further data from these institution data sets.

To achieve this a systematic harmonization of the actors by means of record linkage or data matching methods is required (cf. Christen 2012 and Magerman, Van Looy, and Song 2006). The term "record linkage" means the merging of information from two data sets, which are assumed to refer to the same unit/entity (Herzog, Scheuren, and Winkler 2007:81). These methods are supported by additional lookup tables for deviating notations and actors not included in Amadeus or Research Explorer. The following subsections briefly describe the applied procedure (for the sequence see figure 2).

Phase I: Basic preparations – "basic pre-processing"

In a first step basic preparations for the following harmonization are performed. These include the conversion of the individual data sets into a uniform format (Stata), adjusting for duplicates and adding regional identification numbers like administrative county keys (AGS5) and spatial planning regions (ROR). Usually the starting point for this is the postal code and the place name. In the course of this procedure the individual data records are assigned unique identification numbers with a prefix in accordance with the respective source.

An important task is the decomposition of the author information from the publication data set into its separate components. Due to the fact that in the Web of Science data set at hand the co-authors, together with their address details, are all listed one after another in a single cell, this cell has to be subdivided into individual fields for each author, which subsequently have to be broken down into names, address details, post codes and place of residence. The parser necessary for this task was realized in Stata.

Phase II: Harmonization of records – "pre-cleaning"

In order to standardize the actor names the individual records undergo a pre-cleaning procedure (for the details on this procedure see Ehrenfeld 2015c). The steps basically follow Magerman, Van Looy, and Song (2006). However, they are adjusted to the specific nature of the five data sets used and expanded. The realization of the routines was done in Stata.

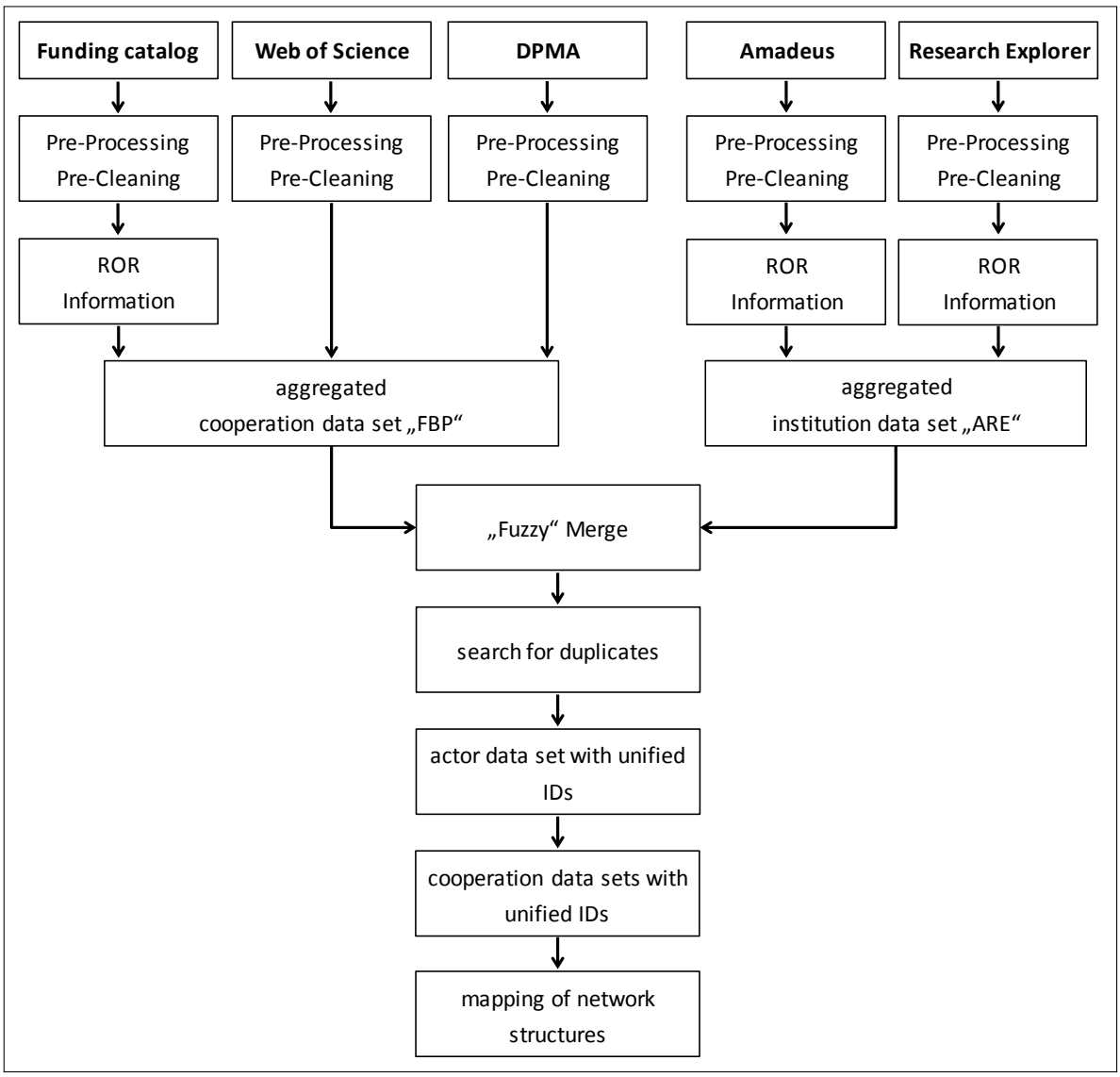

Figure 2: Sequence of the applied data matching procedure.

The first step is a character clean up. For this, the actor name is completely converted to uppercase. It also includes the replacement of German umlauts, accented characters or characters with coding annotations with their ASCII equivalents. Double spaces in the name and spaces at the beginning or end of the name are removed. Subsequently, bracket symbols and notations for "and" are unified and bracketed expressions are extracted.

In a second step all non-ASCII characters are deleted from the name. Then the company structures of business enterprises are identified. This is done through an identification table, which currently holds more than 600 notations for different company structures. The original notations of the corporate structure are subsequently deleted from the company name. Thereafter, notations of frequently used terms with variable spellings are harmonized. Finally, all spaces are removed from the expression (condensation).

These procedures are very suitable to ensure the correct association of actor names with slightly differing spellings. Since parts of this data were collected manually or from scanned paper documents there are occasional errors (like wrong letters). This includes differences in the usage of hyphens and spaces, which prevent a direct association via identity of strings. Failures during the pre-cleaning phase can hardly be compensated for, even through the application of "fuzzy" allocation algorithms. The work put in at this stage is important and a good investment in a reliable allocation.

From a technical point of view the problem of slightly differing notations is very different from the problem of varying denotations for the same institution. Well-known examples for this are Technical Universities (and their frequently used abbreviation "TU") or the "classic" Rheinisch-Westfälische Technische Hochschule Aachen (short: RWTH Aachen). In these cases purely deterministic allocation of original data records or "fuzzy" (probabilistic) methods alone can hardly ensure a reliable allocation. Instead, these cases can be standardized by means of automatized replacement rules or through an additional table for different notations of the same institution.

Phase III: Merging – "data linkage"

At the beginning of this phase the two institution data sets from Amadeus and Research Explorer are merged into an actor reference data set (ARE). The three cooperation data sets (funding catalog, Web of Knowledge, DPMA) are also merged. Subsequently, the two aggregated data sets are merged through the use of probabilistic methods in the form of the commercial software "Fuzzy Dupes"[2]. Compared to purely deterministic procedures, probabilistic methods have the advantage that even non-identical expressions can be linked. For this, however, thresholds have to be defined, and the result should be checked for false positives.

In the following step duplicate names and region features in the cooperation data set are identified in order to group entries for patents, publications and funded projects for the same actor. This step is also done using Fuzzy Dupes. The actor ID, which was assigned to each record during phase I, is replaced by a group ID in case of a successful grouping. This ID is used if an actor can successfully be linked to the actor reference data set. The allocation is checked again in order to minimize false or missing assignments during an iterative procedure.

1.3. Result of the data matching procedure

Following the merger of the cooperation data set with the reference institutions data set, a total of 2810 different actors can be identified. 938 of these can conclusively be identified

[2] Alternatives to this software are the Stata-package "reclink" as well as the free software "MergeToolbox"(Schnell, Bachteler, and Reiher 2005 or Schnell, Bachteler, and Bender 2004). "Fuzzy Dupes" was used due to technical reasons and reasons of speed. For this, a necessary file conversion was accepted.

as business related, 162 can be categorized as research institutes. Further 138 actors can be found in at least two of the cooperation data sets but cannot be conclusively assigned to a record in either the business or the science data set. Therefore, they are grouped and assigned an unique group ID.

At this point there are 1572 unclassified actors, of which 1104 can be identified as natural or private persons from the patent data set. The remaining 468 actors are mainly branches, subordinate locations or facilities of otherwise recorded companies and research institutes, as well as closed companies or institutions (see figure 3). A deeper analysis of these results can be found in Titze et al. (2015).

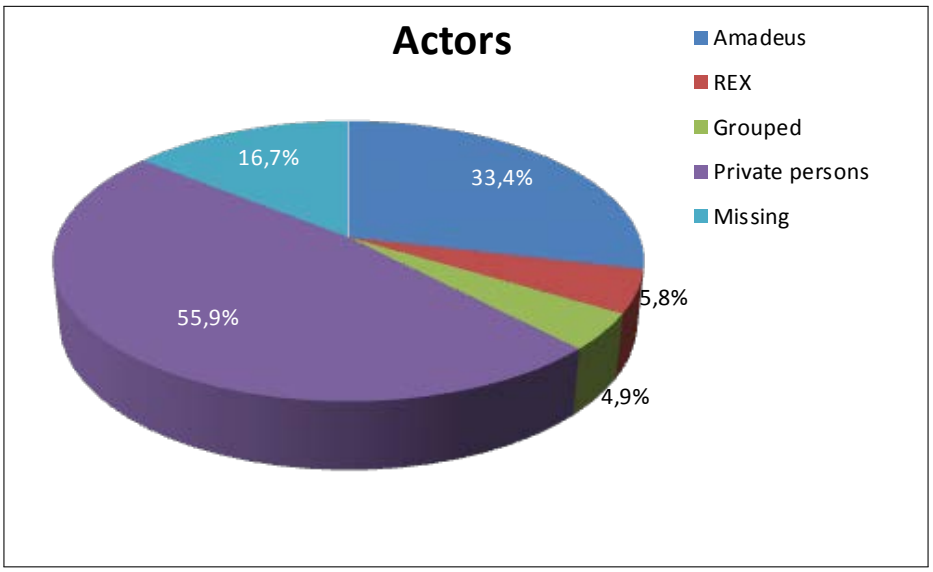

Figure 3: Result of the data matching process: allocation of the actors.

2. Starting basis

The following describes the data sets and procedures in detail from a technical point of view. Primary objective of the presented routines is the mapping of cooperative relations of companies, universities, non-university research facilities and other institutions on three levels of innovation activity (joint projects; publications, patents). In order to do this the innovation-related databases (see below) are linked with the data from the institution data sets Amadeus and Research Explorer by means of record linkage procedures. The structure and sequence of the routines is depicted in figure 2 (p. 6).

The starting basis comprises two institution databases (Amadeus and Research Explorer) and three innovation-related data sets (funding catalog; Web of Knowledge; DPMA German Patents):

1. Individual files company database **Amadeus** 2003 – 2014.
   Data format: Tab separated values (tsv).
   Variables: Mark; Company name; BvD ID number; Zip code; City.

2. Tables database **Research Explorer**.
   Data format: Excel xlsx.
   Overview of German universities and non-university research institutes. "Small version": `20140321_Forschungseinrichtungen_REX.xlsx`; "Large version" (including individual institutes and professorial chairs): `0140507_Forschungseinrichtungen_REX.xlsx`.
   Variables: Id; Institution; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion.

3. Extract from the **BMBF Förderkatalog**.
   Data format: Pipe (|) separated values (psv).
   Filtered for projects with at least one actor from the six RegDemo case regions (cf. appendix A.4).
   Content: Project funding measures, research and development contracts.
   Variables: (see also overview in appendix A.1.): V_nr; FKZ; V_disref; V_ressort; V_thema; V_foeart; V_pt; V_proref; Fi_von; Fi_ende; Fi_sumbew; lpsys; lptext; Name_ze; Ort_Ze; Bundesland_ze; Gemkz_ZE; Gembz_Ze; Land_ZE; Name_St; Ort_St; Bundesland_St; Gemkz_St; Gembz_St; Land_St; Hist_prof; Hist_prof_Klar; Wahlkreis; Wzweig; Ver_nr_vb; Ste_key_ZE; Ste_key_St; V_Stesys.

4. Extract from **Web of Knowledge** or *Web of Science*.
   Data format: Excel xlsx.
   Content: Data on publications ($\rightarrow$ co-publications). Filtered for period 2000-2012. Filtered for the six RegDemo case regions (see below) – only co-authors in case regions. All "external" authors were removed.

   Actors: `ID_ROR_7.xlsx`.
   Variables: Einrichtungen; ID; Stadt; PLZ; Ost/West; Typ; Rechtsform; ROR; AGS 8; AGS 5; Anmerkungen.

   Publications: `Gesamtpuplikationen_ROR_innerhalb_ohneSingle_ohneIntern_2.xlsx`. Individual spreadsheets for Aachen; Dresden; Magdeburg; Kassel; Rostock; Siegen. *Variables:* Pubnr.; ROR Akteure pro Publ.; Akteure pro Pub.; PY; Interne Akteur?; ID; Akteursname.

5. Extract from **DPMA** data.
   Individual files for each of the six RegDemo spatial planning regions. Data format: txt/xlsx mixed.
   Variables: mainid; patnr_dpma; pa; extern; typassignee; orig_inv; pacity; plz; kgs; ror; bl; pacountry.

3. Processing of the institutions databases

3.1. Processing Amadeus

This section describes the processing of the Amadeus data.

01 Converter.do

These 12 routines (the data sets are too different in regard to their structure, which prevents the application of a loop) read the Amadeus data for each year from the individual tsv-files, set variable names and data types, create a variable for the respective year (last_year) and save the files in Stata format. The BvDid comprises 12 characters, the first two of which represent the country code (DE).

Input: `Amadeus_YYYY.tsv`
Mark; Company name; BvD ID number; Zip code; City

Output: `Amadeus_YYYY.dta`

BvDid	Bureau van Dijk ID. ID to identify companies
Name	company name
PLZ	postal code
Ort	place
last_year	year of the data set (2003..2014)

02 Append.do

In this step the individual year files are brought together using the append-function, creating a single combined data set. During this process duplicates in the variables BvDid, Name, PLZ and Ort are identified and removed. The latest version of each record is retained.

Input: `Amadeus_YYYY.dta`
BvDid; Name; PLZ; Ort; last_year

Output: `Amadeus_2003-2014_raw.dta`
BvDid; Name; PLZ; Ort; last_year

03 RLPC Amadeus_2003-2014.do

This routine carries out the record linkage pre cleaning (RLPC) for company names (Name) in the Amadeus data set (for details see Ehrenfeld 2015c). The procedure converts special characters to their ASCII equivalent, identifies company structures and isolates bracketed expressions. This creates an expression which enables a comparison with other company names (RLName). In the course of this process variable names are adapted to the data set (prefix BvD - see appendix A.3). The term for the last year (last_year) is shortened to _lyr.

Input: `Amadeus_2003-2014_raw.dta`
BvDid; Name; PLZ; Ort; last_year

Output: `Amadeus_2003-2014_RLPC.dta`
BvD_ID; BvD_Name; BvD_PLZ; BvD_Ort; BvD_lyr;

RLName	adjusted and modified company name
BvD_legform	identified corporate structure of the company (GmbH, KG, ...)
BvD_temp_name	RLName from last step
BvD_clean_hist	states the numbers of steps the RLName has run through
BvD_brackets	bracketed expression isolated from RLName

04 AGS zuspielen.do

In this step the county code (AGS5) is generated from the postal code or place name and added to the data set. Furthermore, temporary files of the RLPC are deleted (BvD_temp_name; BvD_clean_hist; BvD_brackets).

Input: `Amadeus_2003-2014_RLPC.dta`
BvD_ID; BvD_Name; BvD_PLZ; BvD_Ort; BvD_lyr; RLName; BvD_legform; BvD_temp_name; BvD_clean_hist; BvD_brackets

Output: `Amadeus_2003-2014_AGS.dta`
BvD_ID; BvD_Name; BvD_PLZ; BvD_Ort; BvD_legform; BvD_lyr; RLName;

BvD_AGS5	county code (AGS5)

05 Remove_dup_Name_PLZ_Ort.do

Duplicate data is removed in two steps. The first step adjusts different notations in BvD_Name (RLName - which is identical in the adjusted data sets - will be used for comparing data sets later on in the process). Of the affected records only the latest one or the one with the longest BvD_ID is retained. Duplicates are identified by BvD_ID; RLName; BvD_Ort; BvD_AGS5; BvD_legform. The second step eliminates further differences in the BvD_ID for the same name, place, AGS5 and corporate structure. This is valid here since the merging later on is done using the name (and other properties, but not the BvD_ID).

Input: `Amadeus_2003-2014_AGS.dta`
BvD_ID; BvD_Name; BvD_PLZ; BvD_Ort; BvD_AGS5; BvD_legform; BvD_lyr; RLName

Output: `Amadeus_2003-2014_unified.dta`
BvD_ID; BvD_Name; BvD_PLZ; BvD_Ort; BvD_AGS5; BvD_legform; BvD_lyr; RLName

This data set is merged with the Research Explorer data in step 3.2 (`07 Merge\_REX\_Amadeus.do`).

3.2. Processing Research Explorer

This process describes the processing of the Research Explorer data within the framework of the project RegDemo. A more comprehensive and recent description of the processing of Research Explorer data can be found in Ehrenfeld (2015b).

01a Converter_20140321.do

Data from the "small" version of the Research Explorer is read from the xlsx file. The existing ID field is converted to an ID compatible with the BvD format (REXid). The ID comprises 12 characters, the first 2-3 of which identify the source. Here the source identifier is REX. The following 9 characters are identical to the number from the original ID field. The individual variables are adjusted and formatted. The field Institution is adjusted for disruptive control characters (CR; LF) (Institution1).

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; Institution1; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

01b Converter_20140507.do

This routine performs the same tasks as `01a Converter_20140321.do` for the "large" Research Explorer data set. The only difference is that the Institution field is renamed to Institution 2, so that the data sets can be merged later on.

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; Institution2; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

01c Converter_Additionals.do

Due to the fact that different notations for institutions (for example Rheinisch-Westfälische Technische Hochschule Aachen vs. RWTH Aachen) have been noticed in the process of merging (described below), a table has been created and implemented at this stage in order to allocate these differing notations to the same REX-ID. The tasks in this routine

comprise the reading in and converting of this xlsx table. The functional scope is identical to 01a Converter_20140321.do.

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

Output: Forschungseinrichtungen_Add.dta
REXid; Institution1; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion

02 Append Additionals.do

This step combines the "small" version of the Research Explorer with the additionally compiled notations.

Input: 20140321_Forschungseinrichtungen_REX.dta
Forschungseinrichtungen_Add.dta

Output: Forschungseinrichtungen_REX.dta

03 Prepare_Institute_List.do

Variables not required for the merger are removed (Strasse; Hausnummer; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion). The actor types are transferred to an equivalent numerical statement (see appendix A.5) and equipped with a label matching the original notation.

Input: Forschungseinrichtungen_REX.dta

Output: Institutes_Init.dta
REXid; Institution1; PLZ; Ortsname; Einrichtungstyp; len

04 AGS zuspielen.do

County codes (AGS5) and federal state ("Bundesland" - BuLa) are added to the institutions. Institution1 is renamed to Name.

Input: Institutes_Init.dta

Output: REX_Institutes_AGS.dta
REXid; Name; PLZ; Ort; AGS5; BuLa; Einrichtungstyp

05 RLPC_ResearchExplorer.do

The institutions' names are adjusted for non-ASCII special characters (similar to the names of companies) and a record linkage compatible name is created (RLPC). Corporate structures are identified; bracketed expressions separated. In the course of this process variable names are adapted to the data set (prefix REX - see appendix A.3).

Input: `REX_Institutes_AGS.dta`

Output: `REX_Institutes_AGS_RLPC.dta`
RLName; REX_ID; REX_Name; REX_PLZ; REX_Ort; REX_AGS5; REX_legform; REX_Typ; REX_temp_name; REX_clean_hist; REX_brackets

06 Prepare_REX_Merge.do

The data set is prepared for the merger with Amadeus. This includes the capitalization of all institution and place names (umlauts are maintained and capitalized; "ß" remains unchanged - just as for Amadeus). Different notations of REX_Name are removed, with the exception of the one with the highest REX_ID. Parts of the RLPC no longer required are deleted.

Input: `REX_Institutes_AGS_RLPC.dta`

Output: `REX_Institutes_PreMerge.dta`
RLName; REX_ID; REX_Name; REX_PLZ; REX_Ort; REX_AGS5; REX_legform; REX_Typ

07 Merge_REX_Amadeus.do

The Research Explorer data set is merged with the Amadeus data set (append). Variable names are adjusted to the combined data set (prefix ARE for Amadeus/Research Explorer). Duplicates in the variables RLName; ARE_Ort; ARE_AGS5 are deleted. Type (1 = business) is added to Amadeus records. Blocking variables for the federal state and the first letter of the name are created (block_name; block_ags). The resulting file is saved in the Amadeus data folder.

Input: `REX_Institutes_PreMerge.dta`
`Amadeus_2003-2014_unified`

Output: `Amadeus_2003-2014_REX.dta`
ARE_ID; ARE_Name; ARE_PLZ; ARE_Ort; ARE_AGS5; ARE_Typ; ARE_legform; BvD_lyr; RLName; block_name; block_ags

08 ROR zuspielen.do

Information on spatial planning regions (ROR - see appendix A.4) is added to the merged ARE data. The (external) table used is `ROR2011`. Subsequently, a line number (ARE_line) is created, which might be required later on for a fuzzy matching routine (The Stata routine "reclink" is an example for such a routine - however, in the end "reclink" was not used after all). The csv version of this table will be required later on for the utilization of "Fuzzy Dupes". Appendix A.2 lists the resulting Stata data types and formatting.

Input: `Amadeus_2003-2014_REX.dta`

Output: `ARE_2003-2014.dta`
`ARE_2003-2014.csv`
ARE_line; ARE_ID; ARE_Name; ARE_PLZ; ARE_Ort; ARE_AGS5; ARE_ROR; ARE_Typ; ARE_legform; BvD_lyr; RLName; block_name; block_ags

In a second step all records from regions other than the six case regions are deleted. The line number (ARE_line) is updated.

Output: `ARE_2003-2014 Fallregionen.dta`
`ARE_2003-2014 Fallregionen.csv`
ARE_line; ARE_ID; ARE_Name; ARE_PLZ; ARE_Ort; ARE_AGS5; ARE_ROR; ARE_Typ; ARE_legform; BvD_lyr; RLName; block_name; block_ags

09 Export Excel Sheets.do

This (optional) routine splits the ARE data set into segments, one for each initial letter, and saves the data sheet by sheet as Excel tables. This might be helpful for additional manual research.

Input: `ARE_2003-2014.dta`

Output: `ARE_2003-2014.dta`
ARE_ID; ARE_Name; ARE_Ort; ARE_AGS5; ARE_PLZ

The data set is used during the merger (step 5.1 - `03 Merge ARE IDs.do`).

4. Processing of the innovation related data sets

4.1. Processing funding catalog

The set-up of the funding catalog comprises an ID for each joint project (Ver_nr_vb; later on FK_VbNr – as group ID), as well as a process number for each actor involved (V_nr; later on FK_VNr). Please note that the process number refers to the process - not the actor. Consequently, V_nr does not represent an actor ID.

In the course of preliminary work the data was already read from a txt file; variable labels were allocated and information on the year of the start and the end were created (year_begin; year_end).

Furthermore, outdated information on district codes were updated (territorial reform 2011) and data on the spatial planning regions was added. Subsequently, the data was filtered: a) only projects close to R & D; b) only data with details on ROR; c) years 1991-2010; d) only joint projects. Finally, only projects with at least one actor from one of the six RegDemo case regions are retained. In this case the complete data set was maintained (`Foeka_Fallregionen_WE.dta`).

01 Foeka_Fallregionen_Institutionen_Prepare_RL.do

This routine prepares the funding catalog data for the RLPC routine. For this, the variable names are adjusted to the data set (prefix FK for **F**örder-**K**atalog); capital letters are used for the names of the implementing entity (Name_St); the place of the implementing entity (Ort_St); as well as the names and places of the grant recipients (Name_ZE; Ort_ZE). For the subsequent merger, however, the implementing entity is decisive.

The actor type (FK_Typ) is converted to a numerical format and labeled (see appendix A.5). Furthermore, the following filters are used sequentially: a) only actors from the case regions (FK_FR - at this point all external actors that cooperated with internal actors are dropped); b) only the years 2000-2012. In order to maintain the names of the top level of the institution the name of the entity (FK_Name_St) is split up and - if the expression is too short - expanded for another part of the entity name.

Input: `Foeka_Fallregionen_WE`
V_nr; FKZ; V_disref; V_ressort; V_thema; V_foeart; V_pt; V_proref; Fi_von; Fi_ende; Fi_sumbew; Ipsys; Iptext; Name_ze; Ort_Ze; Bundesland_ze; Gemkz_ZE; Gembz_Ze; Land_ZE; Name_St; Ort_St; Bundesland_St; Gemkz_St; Gembz_St; Land_St; Hist_prof; Hist_prof_Klar; Wahlkreis; Wzweig; Ver_nr_vb; Ste_key_ZE; Ste_key_St; V_Stesys

Output: `Foeka_Fallregionen_Vorhaben.dta`
FK_VbNr; FK_VNr; FK_Name; FK_Name_St2; FK_Ort; FK_AGS5; FK_ROR; FK_FR; FK_Typ; FK_ZE_Name; FK_ZE_Ort; FK_ZE_AGS; year_begin; year_end; FK_Name_St

In a second step the actors are isolated on the branch level. Here the faculty, for example, is still relevant for the distinction between actors.

Output: `Foeka_Fallregionen_Stellen.dta`
FK_Name; FK_Name_St2; FK_Ort; FK_AGS5; FK_ROR; FK_FR; FK_Typ

In the third step, only the top level of the institution is retained. Furthermore, an actor ID is assigned for this level (FKA_ID). The IDs have 12 characters; are consecutively numbered (sorting FK_Name; FK_Ort; FK_AGS5) and have the prefix FKA (**F**örder-**K**atalog-**A**kteur).

Output: `Foeka_Fallregionen_Institutionen.dta`
FKA_ID; FK_Name; FK_Ort; FK_AGS5; FK_ROR; FK_FR; FK_Typ

02 Foeka_Fallregionen_Institutionen_RLPC.do

This routine carries out the RLPC routine for the funding catalog institution data.

Input: `Foeka_Fallregionen_Institutionen.dta`

Output: `Foeka_Fallregionen_Institutionen_RLPC.dta`
FKA_ID; FK_Name; FK_Ort; FK_AGS5; FK_ROR; FK_Typ; FK_legform; RLName

03 Foeka_Fallregionen_Vorhaben_FKA.do

In this step the projects are assigned the IDs of the involved actors. The project data come from `Foeka_Fallregionen_Vorhaben.dta` (see step 01). The IDs come from `Foeka_Fallregionen_Institutionen_RLPC` (see step 02). Appendix A.2 shows the resulting Stata data types and formatting as example for the structural expansion of the data sets.

Input: `Foeka_Fallregionen_Vorhaben.dta`

Output: `Foeka_Fallregionen_Vorhaben_FKA.dta`
FK_VbNr; FK_VNr; FKA_ID; FK_Name; FK_Name_St2; FK_Ort; FK_AGS5; FK_ROR; FK_FR; FK_Typ; FK_ZE_Name; FK_ZE_Ort; FK_ZE_AGS; year_begin; year_end; FK_Name_St; FK_legform

In a second step the data are according to the case regions <FR>.

Output: `Foeka_FKA_<FR>`

This data set will be merged with the bibliometric data and patent data in step 5.1 - 01 `Zusammenführen.do`.

4.2. Processing bibliometric data

For every publication there is a uniform publication number (Pubnr., later on WK_PubNr). This is used as group ID.

01 ID_ROR_7_RLPC.do

This routine creates keys for the numeric codes (ID_ROR_7_Cities; ID_ROR_7_Legforms) used in the actor list (ID_ROR) and assigns the entries for the codes in clear text. It reads the actor data from the tsv file; converts the already assigned actor IDs into the uniform format (WKA_ID; prefix WKA for **W**eb-of-**K**nowledge-**A**ctor) and filters out duplicate actor IDs. Furthermore, the entries for the actor type (WK_Typ) are transformed to match the format used in the other data sets. Name and place are - similar to the other data sets - capitalized. Finally, the RLPC routine is applied to the actor name (WK_Name → RLName).

Input: `ID_ROR_7.tsv`
Einrichtungen; ID; Stadt; PLZ; Ost/West; Typ; Rechtsform; ROR; AGS 8; AGS 5; Anmerkungen

Output: `ID_ROR_7_RLPC.dta`
WKA_ID; WK_Name; WK_PLZ; WK_Ort; WK_AGS5; WK_ROR; WK_Typ; WK_legform; RLName

02 Publikationen_WKA.do

The publication data for the individual case regions <FR> are read in. The existing ID is transformed into the uniform format (WKA_ID). The actor data from `ID_ROR_7_RLPC.dta` (from step 01) is added.

Input: `WK_<FR>.tsv`
Pubnr.; ROR Akteure pro Publ.; Akteure pro Pub.; PY; Interne Akteur?; ID; Akteursname

Output: `WK_WKA_<FR>.dta`
WK_PubNr; WKA_ID; WK_Jahr; Name; WK_Name; WK_PLZ; WK_Ort; WK_AGS5; WK_ROR; WK_Typ; WK_legform

This data set is merged with the funding catalog and patent data in step 5.1 - 01 `Zusammenführen.do`.

4.3. Processing patent data

Each patent has a patent number (patnr_dpma; DP_DPMA). Furthermore, the data set has another ID (mainid; DP_Mainid), which has a better coverage compared to the patent number. Therefore it is used as group ID. The data sets are subdivided into the case regions and contain the name of the patent applicant (pa); the applicant type (1 = company; 2 = university; 3 = non-university research facility; 4 = other - see appendix A.5); the original field of the DPMA data set for the applicant (orig_inv; DP_Orig); as well as place name,

district code/AGS5, postal code, ROR, federal state (bl; DP_BuLa) and country (pacountry; DP_Ctry) of the applicant. The variable bl was generated from the ROR.

Additionally, there is information (extern; DP_Extern) on why the record was included. It states whether the applicant is from the respective ROR (external = nein/0), or whether at least one inventor is from the ROR, but not the applicant (external = ja/1). However, the coverage of the place names is a frequently encountered problem.

One problem in the original data is the spatial allocation of the innovation performance. An example for this is that the applications for all of Audi's patents were filed in Ingolstadt, regardless of where the invention was made. However, often times there is a branch where the invention was made. In these cases the name and ROR (and depending on this also the federal state - but none of the other data!) were modified in such a way that the name now includes the specified regional branch (e.g. Audi AG Kassel).

01 Einlesen.do

The patent data is read from the tsv files (the original data was stored in xlsx files - these have been converted into tsv files). The labels for the actor type (DP_Typ) are unified and data for the spatial planning regions is added (DP_ROR; DP_ROR_Bez).

Input: `DP_<FR>.tsv`
mainid; patnr_dpma; pa; extern; typassignee; orig_inv; pacity; kgs; plz; ror; bl; pacountry

Output: `DP_<FR>.dta`
DP_Mainid; DP_Name; DP_PLZ; DP_Ort; DP_AGS5; DP_ROR; DP_ROR_Bez; DP_BuLa; DP_Ctry; DP_Extern; DP_Typ_old; DP_Typ; DP_DPMA; DP_Orig

02 ROR Filter + Akteure.do

Actor and place names are capitalized. Subsequently, only entries with participating actors from the case regions are retained. Then all single entries are deleted (since the cooperative relations are of interest).

Input: `DP_<FR>.dta`

Output: `DP_ROR_Filter_<FR>.dta`
DP_Mainid; DP_Name; DP_PLZ; DP_Ort; DP_AGS5; DP_ROR; DP_ROR_Bez; DP_BuLa; DP_Ctry; DP_Extern; DP_Typ_old; DP_Typ

In a second step from the patent entries only the actors adjusted for duplicates (DP_Name; DP_Ort; DP_AGS5) are retained.

Output: `DP_Akteur_<FR>.dta`
DP_Name; DP_Ort; DP_AGS5; DP_ROR; DP_Typ

The third step combines the individual actor lists of the spatial planning regions to create one central list. The actors are given an ID according to a unified format (prefix DPA for Deutsches Patent Akteur. Sorting: DP_ROR; DP_Name; DP_AGS5).

Output: `DP_Akteure.dta`
DPA_ID; DP_Name; DP_Ort; DP_AGS5; DP_ROR; DP_Typ

03 Akteure RLPC.do

A name suitable for record linkage is created for the actor names (RLName).

Input: `DP_Akteure.dta`

Output: `DP_Akteure_RLPC.dta`
DPA_ID; DP_Name; DP_Ort; DP_AGS5; DP_ROR; DP_Typ; *DP_legform*; *RLName*

04 Patente_DPA.do

The DPA-IDs from step 03 (`DP_Akteure_RLPC.dta`) are added to the patents from the case regions from step 02 (`DP_ROR_Filter_<FR>.dta`).

Input: `DP_ROR_Filter_<FR>.dta`

Output: `DP_DPA_<FR>.dta`
DP_Mainid; DPA_ID; DP_Name; DP_PLZ; DP_Ort; DP_AGS5; DP_ROR; DP_ROR_Bez; DP_BuLa; DP_Ctry; DP_Extern; DP_Typ_old; DP_Typ; DP_legform

This data set is merged with the funding catalog data and the bibliometric data in step 5.1 - `01 Zusammenführen.do`.

5. Construction of the innovation networks

5.1. Merging the data sets

After the individual components have been processed, they are now merged.

01 Zusammenführen.do

In this step the actors from the three innovation related data sets (funding catalog; bibliometric data and patent data) are merged (`Foeka_Fallregionen_Institutionen_RLPC` from section 4.1; `ID_ROR_7_RLPC` from section 4.2 and `DP_Akteure_RLPC` from section 4.3). The Amadeus and Research Explorer data have already been merged in section 3.2 - `07 Merge_REX_Amadeus.do`. The individual ID columns are combined to create a uniform ID column. At this point there are also some modifications made to RLName in order to capture certain notations with additions.

Input: `Foeka_Fallregionen_Institutionen_RLPC`
`ID_ROR_7_RLPC`
`DP_Akteure_RLPC`

Output: `Foeka_Biblio_Patent_Akteure.csv`
`Foeka_Biblio_Patent_Akteure.dta`
ID; (dup); Name; Ort; AGS5; ROR; legform; Typ; source; RLName

02 Fuzzy Dupes

This step comprises two stages, which make use of the external (commercial) software "Fuzzy Dupes". The settings of this program can be found in appendix A.6 and A.7. In the first stage the line numbers from the combined ARE data set are allocated to the corresponding entries from the combined actor list from step 01.

The FuzzyMatchID created in the course of this allocation refers to the line number (!) of the ARE data set; _Line# is the line number of the actor data set; Matching provides a (numeric) matching score [0;1] and represents goodness of fit between the actor data set and the ARE data set.

Input: `Foeka_Biblio_Patent_Akteure.csv`
`ARE_2003-2014 Fallregionen.csv`

Output: `Fuzzy_Dupes_Step_1.csv`
ID; Name; Ort; AGS5; ROR; legform; Typ; source; RLName; _Line#; Matching; FuzzyMatchID

In the second stage duplicate actors are identified and assigned a Fuzzy Dupes-specific group ID. FuzzyDupesID is the newly created group ID. Actors with the same ID are probably identical; Matching, again, represents the matching score; entries to be deleted are indicated with Delete; however, this parameter is not used here.

Input: `Fuzzy_Dupes_Step_1.csv`

Output: `Fuzzy_Dupes_Step_2.csv`
ID; Name; Ort; AGS5; ROR; legform; Typ; source; RLName; _Line#; Matching; FuzzyMatchID; Matching; FuzzyDupesID; Delete

5.2. Unifying the IDs

03 Merge ARE IDs.do

In the first stage the combined actor list is merged with the combined ARE data set ("left join"). The ARE line number is used as key. In the second stage duplicate actors are identified and assigned a group ID. Finally, a uniform ID is assigned (U_ID for "**U**nified **ID**").

Some variables are renamed or created in the course of the merger and the processing of duplicates: The corresponding ARE line number is renamed from FuzzyMatchID into mline (for matching line - will be deleted later on); this is used for, among other things, the merger of ARE_ID; ms indicates the matching score of the actor list and the ARE data set.

The matching score for actor duplicates is ds; cFD counts the entries in a Fuzzy Dupes group. A new group ID is created by grouping Fuzzy Dupes IDs according to the FDG_ID; cml counts (similarly to cFD) the number of entries, which simultaneously have the same FDG_ID and the same matching line (to ARE). Comparing cFD to cml serves the purpose of finding discrepancies between the grouping by means of duplicates and the match to ARE.

The unified ID (U_ID) is assigned hierarchically: If an entry for an ARE-ID exists, this ID is adopted. If there is no ARE, the group ID (FDA_ID) is used. In the case that this also does not exist, the original ID from the actor data set is used instead (FKA/WKA/DPA); cUID counts the entries for each group with the same UID. Group sizes might turn out larger than cFD and cml since the allocation of ARE and the identification of duplicates are done separately but, in the end, lead to the same UID.

Input: `Fuzzy_Dupes_Step_2.csv`

Output: `Akteure_FD_ARE.dta`
U_ID; cUID; FDG_ID; ds; cFD; ARE_ID; ms; cml; ID; Name; ARE_Name; Ort; ARE_Ort; AGS5; ARE_AGS5; ROR; ARE_ROR; legform; ARE_legform; Typ; ARE_Typ; RLName; ARE_RLName; ARE_PLZ

A table containing only the original ID (ID) and the unified ID (U_ID) is created in order to facilitate the addition to the individual innovation data sets.

Output: `Akteure_FD_ARE_UID.dta`
ID; U_ID

04 Vorhaben UIDs.do

In this step the U_IDs are added to the innovation data sets in their respective directories. The original group labels (FK_VbNr; WK_PubNr; DP_Mainid) are transferred to the unified variable Group. The actor IDs are listed as Node; the respective names as Node_Name. Node_nr describes the sequential number of the node within its group.

Input: Akteure_FD_ARE_UID.dta
Foeka_FKA_<FR>.dta
WK_WKA_<FR>.dta
DP_DPA_<FR>.dta

Output: Foeka_UID_<FR>.dta
WK_UID_<FR>.dta
DP_UID_<FR>.dta
Group; Node_nr; Node; Node_Name

5.3. Construction of the network structures

05 Networks.do

This routine creates fully linked, undirected networks from each of the three innovation data sets. The edges connect two nodes each (Node1; Node2).

Input: Foeka_UID_<FR>.dta
WK_UID_<FR>.dta
DP_UID_<FR>.dta

Output: Foeka_Network_<FR>.dta
WK_Network_<FR>.dta
DP_Network_<FR>.dta
Group; Node1_nr; Node2_nr; Node1; Node2; Node1_Name; Node2_Name

The frequencies of the resulting edges are counted in the second stage of this step.

Output: Foeka_Freq_<FR>.dta
WK_Freq_<FR>.dta
DP_Freq_<FR>.dta
Node1; Node2; Freq; Node1_Name; Node2_Name

06 Comparison Table.do

This stage summarizes the frequency distributions of the three innovation data sets for each spatial planning region. They can be used to compare the innovation activity for each edge of the previously created networks.

Input: Foeka_Freq_<FR>.dta
WK_Freq_<FR>.dta
DP_Freq_<FR>.dta

Output: Comparison_<FR>.dta
total; Node1; Node2; FK_Freq; WK_Freq; DP_Freq; Node1_Name; Node2_Name

In a second step the tables for all spatial planning regions are aggregated into a single table.

Output: `Comparison_Total.dta`

References

Christen, Peter (2012): Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Berlin, Heidelberg: Springer.

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.

Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler (2007): Data Quality and Record Linkage Techniques. New York: Springer Science+Business.

Magerman, Tom, Bart Van Looy, and Xiaoyan Song (2006): Data production methods for harmonized patent statistics: Patentee name harmonization. Katholieke Universiteit Leuven MSI 0605.

Schnell, Rainer, Tobias Bachteler, and Stefan Bender (2004): A Toolbox for Record Linkage. In: *Austrian Journal of Statistics* 33.1–2, pp. 125–133.

Schnell, Rainer, Tobias Bachteler, and Jörg Reiher (2005): MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. ZA-Information 2005, No. 56.

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. Variables funding catalog

Name	Description
V_nr	Operation number ("Vorhaben-Nummer") (unique key field)
FKZ	Funding code
V_disref	Unit of the respective department
V_ressort	Department
V_thema	Project topic
V_foeart	Type of funding
V_pt	Project manager
V_proref	Competent entity of the project manager
Fi_von	Start
Fi_ende	End
Fi_sumbew	Total funding amount of project
lpsys	Benefit plan number ("Leistungsplan-Nummer")
lptext	("Leistungsplan-Beschreibung")
Name_ze	Grant recipient
Ort_Ze	Place of grant recipient
Bundesland_ze	Federal state of grant recipient
Gemkz_ZE	District code of place of grant recipient
Gembz_Ze	Name of municipality on district code level to grant recipient
Land_ZE	Country of grant recipient (if not Germany)
Name_St	Implementing entity
Ort_St	Place of implementing entity
Bundesland_St	Federal state of implementing entity
Gemkz_St	District code of place of implementing entity
Gembz_St	Name of municipality on district code level to implementing entity
Land_St	Country of implementing entity (if not Germany)
Hist_prof	Förderprofil
Hist_prof_Klar	Förderprofil-Beschreibung
Wahlkreis	Constituency number (current legislative period)
Wzweig	Classification of economic sector
Ver_nr_vb	Joint project number ("Verbund-Kennzeichen")
Ste_key_ZE	Classification scheme of the grant recipient ("Stellensystematik des Zuwendungsempfängers")
Ste_key_St	Classification scheme of the implementing entity ("Stellensystematik der ausführenden Stelle")
V_Stesys	Classification scheme of the operation ("Stellensystematik auf Vorhabenebene")

A.2. Data types and presentation - general structure

Name	Type	Format	Remarks
Example actors: [ARE_2003-2014.dta]			
ARE_line	float	%9.0g	
ARE_ID	str12	%12s	
ARE_Name	strL	%-100s	
ARE_PLZ	str5	%9s	
ARE_Ort	str39	%-27s	
ARE_AGS5	str5	%9s	
ARE_ROR	int	%10.0g	
ARE_Typ	byte	%-20.0g	
ARE_legform	str17	%-15s	
BvD_lyr	int	%9.0g	
RLName	str165	%-100s	
block_name	str1	%9s	
block_ags	byte	%10.0g	
Example innovation project: [Foeka_Fallregionen_Vorhaben_FKA.dta]			
FK_VbNr	long	%10.0g	group ID
FK_VNr	long	%9.0g	
FKA_ID	str12	%12s	actor ID
WK_Jahr	str4	%9s	
FK_Name	str150	%-95s	
FK_Name_St2	str167	%-75s	
WK_PLZ	str5	%9s	
FK_Ort	str23	%23s	
FK_AGS5	str5	%9s	
FK_ROR	int	%10.0g	
DP_ROR_Bez	str10	%10s	
DP_BuLa	str2	%9s	
DP_Ctry	str2	%9s	
DP_Extern	byte	%10.0g	
FK_FR	byte	%1.0f	
FK_Typ	byte	%37.0g	
FK_ZE_Name	strL	%-75s	
FK_ZE_Ort	str23	%23s	
FK_ZE_AGS	str5	%9s	
year_begin	int	%10.0g	
year_end	int	%10.0g	
FK_Name_St	strL	%-100s	
FK_legform	str17	%-15s	

A.3. Data source in IDs

Format: 12 characters

Prefix	Meaning
DE	Amadeus Germany
REX	Research Explorer
FDA	Actors grouped by means of Fuzzy Dupes
WE	Allocated to groups manually
FKA	Funding catalog actor
WKA	Web of Knowledge actor
DPA	German Patent Office actor

A.4. Case regions RegDemo

ROR	Meaning
501	Aachen
513	Siegen
602	Nordhessen (="Kassel")
1302	Mittleres Mecklenburg/Rostock (="Rostock")
1401	Oberes Elbtal/Osterzgebirge (="Dresden")
1504	Magdeburg

A.5. Coding of actor types

Code	Meaning
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.6. Fuzzy Dupes step 1: Merger with ARE data set

* New Project *

[Database Connection]
Datasource: Text/CSV.
< Open... >

Filename: `Foeka_Biblio_Patent_Akteure.csv`

[CVS Options]
Delimiter: Tab Stop; Text Delimiter: "
Decimal Delimiter: Comma
Column Headlines: check
<Refresh> <OK>

[Special Fields]
<Keine Einträge>

[Duplicate Fields]
ROR
Cluster: check
Dupe Search: check
Weight: Lower
NULL-Compare: check

RLName
Dupe Search: check
Weight: Higher
NULL-Compare: check

< Next -> >

[Normalization]
ROR
Standard: uncheck

RLName
Standard: uncheck

< Next -> >

[Options]
Threshold Cluster: 0,500 (Rechtsanschlag)

Threshold Duplicates: 95%

< OK >
save as:
`g:\Wed\# aktuelle Arbeit\RegDemo\Zusammenführung\Fuzzy_Dupes_Step_1.prj`

* Dupe Search - Match with second list *

[Database Connection]
Datasource: Text/CSV.
< Open... >

g:\Wed\# aktuelle Arbeit\RegDemo\Zusammenführung\ARE_2003-2014.csv
oder
... \ARE_2003-2014 Fallregionen.csv

[CVS Options]
Delimiter: Tab Stop; Text Delimiter: "
Decimal Delimiter: Comma
Column Headlines: check
<Refresh> <OK>

Reading data...

[Fuzzy Match with external list]

Target Field - Values from Import Source
ROR - ARE_ROR
RLName - RLName

< Import >

[Fuzzy Match with Externam List]
[Return Results]
All Records (tagged)

Threshold Cluster: 0,500 (Rechtsanschlag)

Threshold Duplicates: 95%

< OK >

Fuzzy Match. ETA: 1h bzw. 7h

Ergebnis speichern als
g:\Wed\# aktuelle Arbeit\RegDemo\Zusammenführung\Fuzzy_Dupes_Step_1.csv

A.7. Fuzzy Dupes step 2: Identification of duplicates –> group IDs

* New Project *

[Database Connection]
Datasource: Text/CSV.
< Open... >

Filename: `Fuzzy_Dupes_Step_1.csv`

[CVS Options]
Delimiter: Colon; Text Delimiter: "
Decimal Delimiter: Comma
Column Headlines: check
<Refresh> <OK>

[Special Fields]
<Keine Einträge>

[Duplicate Fields]
ROR
Cluster: check
Dupe Search: check
Weight: Lower
NULL-Compare: check

RLName
Dupe Search: check
Weight: Higher
NULL-Compare: check

< Next -> >

[Normalization]
ROR
Standard: uncheck

RLName
Standard: uncheck

< Next -> >

[Options]
Threshold Cluster: 0,500 (Rechtsanschlag)

Threshold Duplicates: 95%

< OK >
save as:
`g:\Wed\# aktuelle Arbeit\RegDemo\Zusammenführung\`
`Fuzzy_Dupes_Step_2.prj`

* Dupe Search - Dupe Search *

[Duplicates Search options]

All records with duplicates tag: check

Threshold Cluster: 0,500 (Rechtsanschlag)

Threshold Duplicates: 95%

< OK >

Duplicate Search. ETA: 2 Minutes

Export - Export
<Optionen lassen wie sie sind>
Colums: Export all columns: check
Rows: Export all rows: check
First line contains column headers: check
Strings in double quotes: check
Delimiter: Colon
Character Set: ANSI (Standard)
<OK>

Ergebnis speichern als
`g:\Wed\# aktuelle Arbeit\RegDemo\Zusammenführung\`
`Fuzzy_Dupes_Step_2.csv`

A.8. Code Statistics

Stand: August 2014

Modul	Anzahl Zeilen
2.1 Aufbereitung Amadeus	
01 Converter.do	55
02 Append.do	86
03 RLPC Amadeus_2003-2014.do	136
04 AGS zuspielen.do	93
05 Remove_dup_Name_PLZ_Ort.do	108
2.2 Aufbereitung Research Explorer	
01a Converter_20140321.do	73
01b Converter_20140507.do	74
01c Converter_Additionals.do	82
02 Append Additionals.do	41
03 Prepare_Institute_List.do	126
04 AGS zuspielen.do	90
05 RLPC_ResearchExplorer.do	137
06 Prepare_REX_Merge.do	91
07 Merge_REX_Amadeus.do	187
08 ROR zuspielen.do	95
09 Export Excel Sheets.do	78
3.1 Aufbereitung Förderkatalog	
01 Foeka_Fallregionen_Institutionen_Prepare_RL.do	346
02 Foeka_Fallregionen_Institutionen_RLPC.do	149
03 Foeka_Fallregionen_Vorhaben_FKA.do	111
3.2 Aufbereitung Bibliometriedaten	
01 ID_ROR_7_RLPC.do	295
02 Publikationen_WKA.do	111
3.3 Aufbereitung Patentdaten	
01 Einlesen.do	184
02 ROR Filter + Akteure.do	148
03 Akteure RLPC.do	101
04 Patente_DPA.do	67
4 Zusammenführung der Datensätze	
01 Zusammenführen.do	166
02 Fuzzy Dupes	extern
03 Merge ARE IDs.do	312
04 Vorhaben UIDs.do	165
05 Networks.do	178
06 Comparison Table.do	115
Anzahl Module	36
Codezeilen Gesamt	**4000**

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