Leibniz-Institut für
Wirtschaftsforschung
Halle

IWH TECHNICAL REPORTS

RegDemo: Aufbereitung und Zusammenführung der Akteursdaten

Technische Dokumentation der Routinen und Datensätze

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: Aufbereitung und Zusammenführung der Akteursdaten – Technische Dokumentation der Routinen und Datensätze. IWH Technical Reports 01/2015. Halle (Saale) 2015.

ISSN 2365-9076

RegDemo: Aufbereitung und Zusammenführung der Akteursdaten

Technische Dokumentation der Routinen und Datensätze

Zusammenfassung

Übergeordnetes Ziel der hier vorgestellten Routinen ist die Abbildung von Kooperationsbeziehungen von Unternehmen, Universitäten, außeruniversitären Forschungseinrichtung sowie sonstiger Institutionen auf drei Ebenen von Innovationstätigkeit (Verbundprojekte; Publikationen; Patente). Dazu werden a) die drei innovationsbezogenen Datenbanken (Förderkatalog; Web of Knowledge; DPMA Patente) vereinheitlicht und zusammengeführt sowie b) diese kombinierten Datenbestände mittels Record-Linkage-Techniken mit den Daten aus den Institutionendatensätzen Amadeus und Research Explorer verknüpft. Die Zusammenführung erstreckt sich für dieses Projekt auf die sechs Fallregionen, die in RegDemo vorgesehen sind. Die Abgrenzung erfolgt anhand von Raumordnungsregionen: 501 - Aachen; 513 - Siegen; 602 - Nordhessen (= „Kassel“); 1302 - Mittleres Mecklenburg/Rostock (= „Rostock“); 1401 - Oberes Elbtal/Osterzgebirge (= „Dresden“); 1504 - Magdeburg.

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Problemstellung und Beschreibung des Verfahrens

Eine Beschränkung auf nur eine Art von Kooperationsbeziehungen wie beispielsweise Patente, Publikationen oder gemeinsam eingeworbene Drittmittelprojekte wird der realistischen Abbildung von Kooperationsbeziehungen zwischen Akteuren nicht gerecht. Zur Minderung dieses Problems wurde daher im Projekt „RegDemo“ ein Mehrebenenansatz entwickelt, der die Kooperationsbeziehungen der Akteure auf den einzelnen Ebenen explizit berücksichtigt und diese Ebenen vereinigt (siehe Titze et al. 2015). Dazu ist es notwendig, die verschiedenen innovationsbezogenen bzw. kooperationsbezogenen Datensätze systematisch zu harmonisieren und zu vereinigen.

Im Folgenden werden die verwendeten Datensätze und das gewählte Verfahren beschrieben. Die Datenbasis der Analyse besteht aus zwei Institutions-Datenbanken (Amadeus und Research Explorer) sowie drei kooperationsbezogenen Datensätzen (Förderkatalog; Web of Science; Deutsche Patente - siehe Abbildung 1). Das verwendete Verfahren basiert auf den Grundlagen von Record-Linkage-Methoden (siehe hierzu beispielsweise Christen 2012 sowie Magerman, Van Looy und Song 2006).

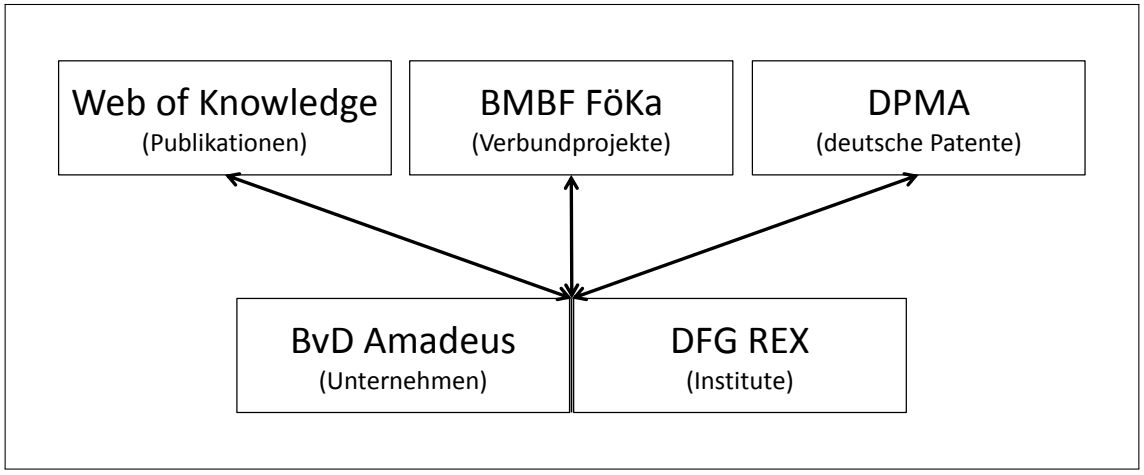

Abbildung 1: Übersicht über die für die Analyse verknüpften Datensätze.

1.1. Beschreibung der einzelnen Datensätze

Amadeus

Die Amadeus-Datenbank des kommerziellen Anbieters Bureau van Dijk umfasst Informationen zu Wirtschaftsunternehmen in Europa mit aktuell 14 Mio. Einträgen. Ihr Inhalt umfasst Kenngrößen aus der Bilanz sowie Gewinn- und Verlustrechnung, weitere Finanzkennzahlen und Angaben zu Konzernstrukturen für die enthaltenen Unternehmen. Für Deutschland sind ca. 1,5 Mio. Unternehmen enthalten. Jedes Unternehmen in diesem Datensatz trägt eine eindeutige Identifikationsnummer (BvD-ID), die später zur Identifikation und Verknüpfung der Akteure verwendet wird. Um sowohl verschiedene in der Vergangenheit verwendete

Schreibweisen für ein existierendes Unternehmen zu erfassen, als auch nicht mehr bestehende Unternehmen identifizieren zu können, wurde ein historischer Amadeus-Datensatz aufgebaut. Dieser umfasst ca. 2,9 Mio. Einträge für den Zeitraum 2003-2014. Um im späteren Verlauf der Prozedur eine schnellere Verarbeitung zu erreichen, werden nur die Datensätze für die sechs Fallregionen in dieser Analyse verwendet. Der verbleibende Datensatz umfasst noch 113.574 Einträge. Durch diesen Eingriff laufen die verwendeten Prozeduren deutlich schneller ab. Eine Beeinträchtigung der Zuordnungsqualität wurde nicht beobachtet.

Research Explorer

Der Research Explorer (kurz: REX) wird kostenlos als Online-Datenbank[1] bereitgestellt. Dieses Verzeichnis umfasst ca. 23.000 Einträge zu Instituten an deutschen Hochschulen sowie außeruniversitäre Forschungseinrichtungen. Bei Universitäten ist die Untergliederung dabei bis auf Lehrstuhlebene abgebildet. Die Angaben sind dabei nach geografischen, fachlichen und strukturellen Kriterien geordnet. Diese Datenbank ergänzt die oben beschriebene Unternehmensdatenbank im Bezug auf Kooperationsakteure, da in dieser für gewöhnlich keine Hochschulen und Forschungseinrichtungen erfasst werden.

Förderkatalog

Der Förderkatalog des Bundesministeriums für Bildung und Forschung (BMBF) listet Daten zu knapp 99.000 laufenden und abgeschlossenen von deutschen Bundesministerien geförderten Forschungsvorhaben. Unter den geförderten Projekten sind knapp 11.000 Verbundprojekte für Deutschland. Für die sechs Fallregionen sind 2.416 Datensätze für 914 Verbundprojekte verzeichnet. Zu den erfassten Angaben gehören die geförderte Stelle, zeitlicher Rahmen, Titel und thematischer Bezug des Projekts. Verbundprojekte tragen eine eindeutige Verbundnummer. Jeder einzelne Fördervorgang eines Projektes ist mit einer eigenen Vorgangsnummer erfasst. Für diese Analyse betrachten wir alle Förderprojekte in den Fallregionen von 1991 bis 2010, die einen Bezug zu Forschung und Entwicklung aufweisen.

DPMA Datenbank

Die Datenbank des deutschen Patent- und Markenamtes (DPMA) enthält Rahmendaten zu deutschen Patenten. Diese Angaben umfassen neben einer eindeutigen Patentnummer und dem Titel des Patents die Namen sowie regionale Informationen zu Erfindern als auch Anmeldern. Zur Analyse werden 1.371 Co-Patente im Zeitraum (Anmeldedatum) von 1994 bis 2010 betrachtet, von denen jeweils mindestens ein Anmelder oder Erfinder in den Fallregionen liegt und mindestens eine Kooperationsbeziehung innerhalb der jeweils betrachteten Fallregion existiert.

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

Web of Science

Die Online-Zitationsdatenbank Web of Science bzw. ehemals ISI Web of Knowledge wird heute vom kommerziellen Anbieter Thomson Reuters betrieben. Zur Analyse standen folgende Pakete dieser Datenbank zur Verfügung: Science Citation Index Expanded (SCI-EXPANDED); Arts & Humanities Citation Index (A&HCI); Social Sciences Citation Index (SSCI). Die ausgewerteten Daten für die Co-Publikationen innerhalb der Fallregionen umfassen den Zeitraum 2000 bis 2012 und beinhalten die Daten für 12.502 Publikationen.

1.2. Vorgehensweise

Ziel der Prozedur ist a) die Zusammenführung der drei Kooperationsdatensätze (Förderkatalog; Web of Science; DPMA Patente) und b) die Verknüpfung dieses kombinierten Datenbestandes mit den Institutionsdatensätzen Amadeus und Research Explorer zur eindeutigen Identifizierung der Akteure. Dieses Verfahren wird gewählt um einerseits einen Normalisierungsstandard für die Schreibweisen der Akteure zu erzeugen und andererseits um weitere Daten aus diesen Institutionsdatensätzen hinzuziehen zu können.

Hierzu ist eine systematische Harmonisierung der Akteure unter Nutzung von Record-Linkage- bzw. Data-Matching Techniken nötig (vgl. Christen 2012 sowie Magerman, Van Looy und Song 2006). Der Begriff „Record Linkage“ bezeichnet dabei das Zusammenführen von Informationen zweier Datensätze, von denen angenommen wird, dass sie sich auf dieselbe Einheit/Entität beziehen (Herzog, Scheuren und Winkler 2007:81). Diese Methoden werden unterstützt, indem zusätzlich ergänzende Lookup-Tables für abweichende Schreibweisen sowie nicht in Amadeus oder Research Explorer erfasste Akteure verwendet werden. Im Folgenden wird das angewandte Verfahren kurz beschrieben (zum Ablauf siehe Abbildung 2).

Phase I: Grundlegende Vorbereitungen – „Basic Pre-Processing“

In einem ersten Schritt werden grundlegende Vorarbeiten zur anschließenden Harmonisierung geleistet. Dazu gehören die Konvertierung der einzelnen Datensätze in ein einheitliches Format (Stata), Bereinigung um Duplikate sowie das Zuspielen regionaler Identifikationsnummern wie Kreisschlüssel (AGS5) und Raumordnungsregion (ROR). Ausgangspunkt ist hierbei in der Regel die Postleitzahl und der Ortsname. Im Zuge dessen werden den einzelnen Datensätzen eindeutige Identifikationsnummern mit einem ihrer Quelle entsprechenden Präfix zugeordnet.

Eine der zentralen Aufgaben hierbei ist das Zerlegen der Autorenangaben im Publikationsdatensatz in seine einzelnen Bestandteile. Da im vorliegenden Web of Science Datensatz alle Koautoren mit ihren Adressangaben nacheinander in einem Feld stehen, muss dieses Feld in einzelne Autorenfelder und anschließend in Namen, Adressen, Postleitzahlen und Wohnorte aufgespalten werden. Der für diese Aufgabe notwendige Parser wurde in Stata realisiert.

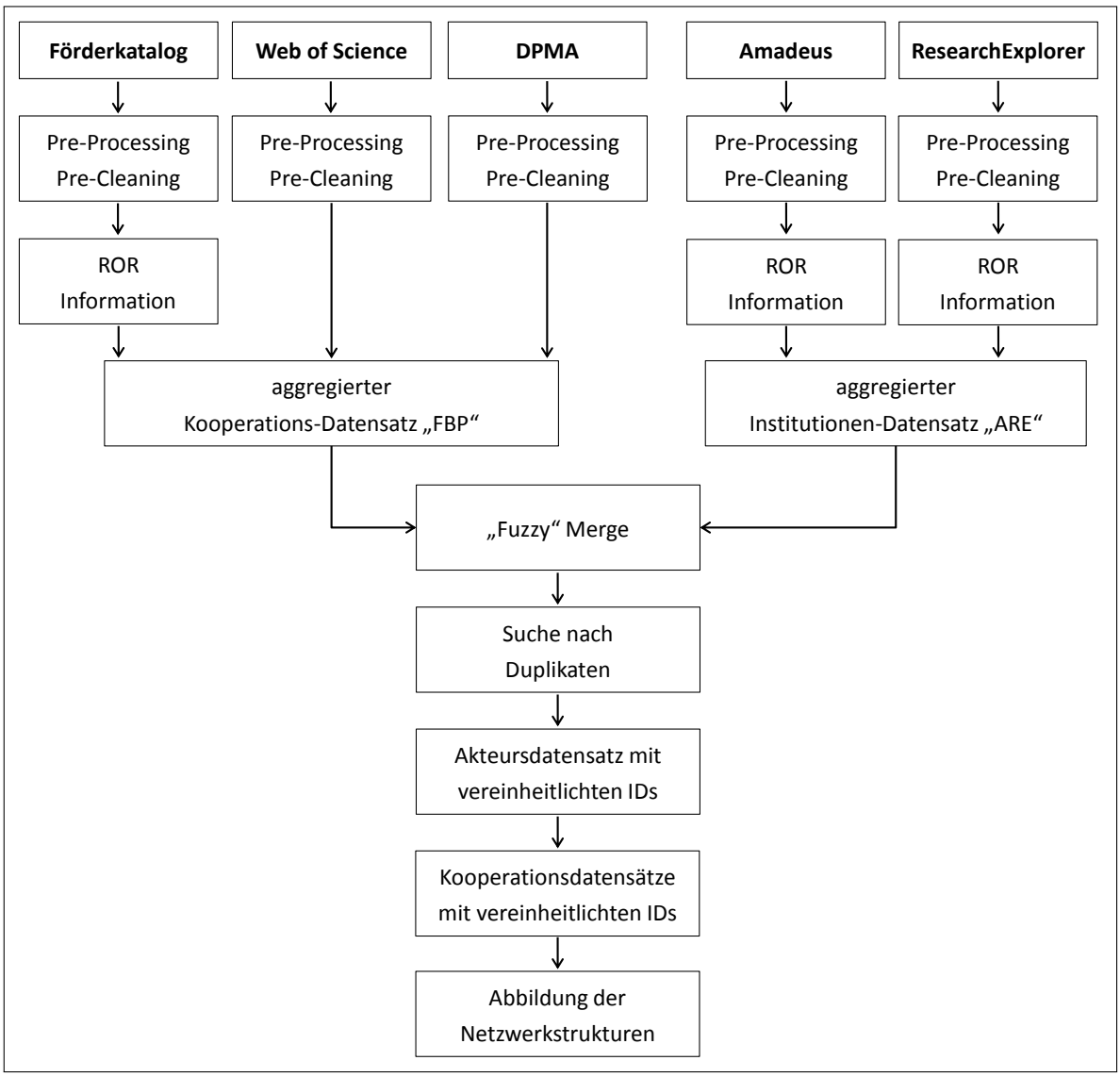

Abbildung 2: Ablauf des verwendeten Data-Matching-Verfahrens.

Phase II: Harmonisierung der Datensätze – „Pre-Cleaning“

Zur Standardisierung der Akteursnamen werden die einzelnen Datensätze einer Pre-Cleaning Prozedur unterzogen (zu den Details dieses Verfahrens siehe Ehrenfeld 2015c). Die einzelnen Schritte folgen im Grunde Magerman, Van Looy und Song (2006), werden aber für die Besonderheit der fünf verwendeten Datensätze entsprechend angepasst und erweitert. Die Realisierung der Routinen erfolgte in Stata.

Zuerst erfolgt eine Zeichenbereinigung. Dazu wird der Akteursname komplett in Großbuchstaben umgewandelt. Im Zuge dessen werden auch deutsche Umlaute, Zeichen mit Akzenten oder Zeichen mit Codierungsvermerken durch ihre ASCII-Äquivalente ersetzt. Doppelte Leerzeichen im Namen sowie Leerzeichen am Beginn oder Ende des Namens werden entfernt. Anschließend werden Klammersymbole und Schreibweisen für „und“ vereinheitlicht sowie vorhandene Klammerausdrücke extrahiert.

In einem zweiten Schritt werden alle Nicht-ASCII-Zeichen aus dem Namen entfernt. Anschließend werden die Gesellschaftsformen von Unternehmen identifiziert. Dies geschieht mittels einer Identifikationstabelle, die zum aktuellen Zeitpunkt über 600 Schreibweisen für verschiedene Gesellschaftsformen umfasst. Die Originalschreibweisen der Gesellschaftsform werden anschließend aus dem Unternehmensnamen entfernt. Schließlich werden einige Schreibweisen von häufig verwendeten, aber unterschiedlich geschriebenen Begriffen harmonisiert. Zuletzt werden alle Leerzeichen aus dem Ausdruck entfernt (Kondensierung).

Diese Prozeduren eignen sich sehr gut um eine Zuordnung bei leicht abweichenden Schreibweisen für denselben Akteursnamen zu gewährleisten. Da Teile dieser Daten manuell erfasst wurden oder aus Scans von Dokumenten in Papierform stammen, sind auch einzelne Buchstabenfehler in den Datensätzen vorhanden. Hierzu gehören auch Unterschiede in der Verwendung von Trennstrichen oder Leerzeichen, die eine direkte Zuordnung mittels Identität von Strings verhindern. Versäumnisse in der Phase des Pre-Cleanings können selbst durch die Verwendung „unscharfer“ Zuordnungsalgorithmen kaum ausgeglichen werden. Die hier geleistete Arbeit ist wichtig und eine gute Investition in eine sichere Zuordnung.

Von der Situation leicht abweichender Schreibweisen zu unterscheiden sind im technischen Sinne deutlich andere Bezeichnungen für dieselbe Institution. Als Beispiel hierfür seien die Technischen Universitäten genannt (mit ihrer oft verwendeten Kurzform „TU“) sowie (als „Klassiker“) die Rheinisch-Westfälische Technische Hochschule Aachen (kurz: RWTH Aachen). Mit einer rein deterministischen Zuordnung der Originaldatensätze oder „unscharfer“ (probabilistischer) Methoden alleine ist es in diesen Fällen kaum möglich, eine sichere Zuordnung zu gewährleisten. Diese Fälle können zur Standardisierung mittels automatisierter Ersetzungsregeln bzw. mit einer zusätzlichen Tabelle für unterschiedliche Schreibweisen derselben Institution erfasst werden.

Phase III: Zusammenführung – „Data Linkage“

Zu Beginn dieser Phase werden die beiden Institutionsdatensätze aus Amadeus und Research Explorer zu einem Akteurs-Referenzdatensatz (ARE) vereint. Ebenso werden die drei Kooperationsdatensätze (Förderkatalog, DPMA, Web of Knowledge) vereinigt. Durch die Verwendung probabilistischer Methoden in Form der kommerziellen Software „Fuzzy Dupes“[2] werden die beiden aggregierten Datensätze anschließend zusammengeführt. Gegenüber rein deterministischer Verfahren haben probabilistische Methoden den Vorteil, dass auch nicht vollständig übereinstimmende Ausdrücke zusammengeführt werden können. Hierzu müssen jedoch Schwellenwerte definiert und auf eventuell falsch positive Zuordnungen kontrolliert werden.

Im darauf folgenden Schritt werden Duplikate bzgl. Name und Regionalmerkmal im Kooperationsdatensatz identifiziert, um die Einträge für Patente, Publikationen und geförderte

[2] Alternativen zu dieser Software sind das Stata-Paket „reclink“ sowie die kostenlos zur Verfügung stehende Software „Merge-Toolbox“ (Schnell, Bachteler und Reiher 2005 bzw. Schnell, Bachteler und Bender 2004). Aus technischen Gründen bzw. Gründen der Geschwindigkeit wurde jedoch „Fuzzy Dupes“ verwendet und eine so notwendig gewordene Dateikonvertierung in Kauf genommen.

Projekte für denselben Akteur zu gruppieren. Dieser Schritt wird ebenfalls mit Fuzzy Dupes durchgeführt. Die in Phase I für jeden Datensatz vergebene Akteurs-ID wird bei erfolgreicher Gruppierung durch eine Gruppen-ID ersetzt. Bei erfolgreicher Zuordnung eines Akteurs zum Akteurs-Referenzdatensatz wird diese ID verwendet. Zur Sicherheit wird die Zuordnung einer wiederholten Sichtprüfung unterzogen um in einem iterativen Verfahren falsche oder fehlende Zuordnungen zu minimieren.

1.3. Ergebnis des Data-Matching-Verfahrens

Nach erfolgter Vereinigung des Kooperationsdatensatzes mit dem Referenz-Institutionsdatensatz können insgesamt 2810 unterschiedliche Akteure identifiziert werden. Davon werden 938 eindeutig als der Wirtschaft zugehörig identifiziert, 162 Akteure können eindeutig Forschungseinrichtungen zugeordnet werden. Weitere 138 Akteure werden in mindestens zwei der Kooperationsdatensätze gefunden, können jedoch nicht eindeutig einem Eintrag der Wirtschaft- oder Wissenschaftsdatensätze zugeordnet werden. Diese werden daher gruppiert und mit einer eindeutigen Gruppen-ID versehen.

Von den 1572 zu diesem Zeitpunkt nicht klassifizierten Akteuren können 1104 als natürliche Personen bzw. Privatpersonen identifiziert werden, die aus dem Patent-Datensatz stammen. Bei den verbleibenden 468 Akteuren handelt es sich vor allem um Zweigstellen, abhängige Standorte oder Werke anderweitig erfasster Unternehmen und Forschungseinrichtungen sowie bereits aufgelöste Unternehmen oder Institutionen (siehe Abbildung 3). Die weitere Auswertung dieser Ergebnisse ist in Titze et al. (2015) beschrieben.

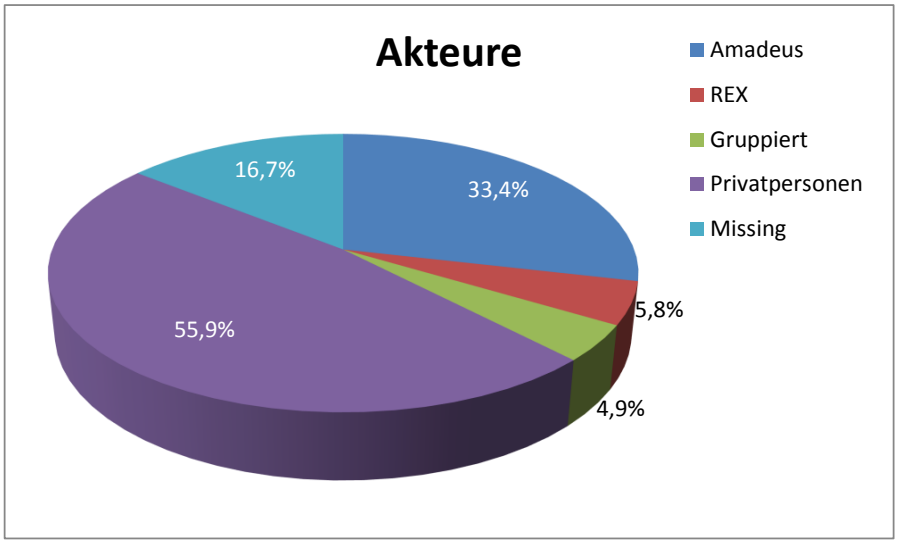

Abbildung 3: Ergebnis des Data-Matching-Prozesses: Zuordnung der Akteure.

2. Ausgangsbasis

Im Folgenden werden die Datensätze und Abläufe aus technischer Sicht detailliert beschrieben. Ziel der hier vorgestellten Routinen ist die Abbildung von Kooperationsbeziehungen von Unternehmen, Universitäten, außeruniversitären Forschungseinrichtungen sowie sonstiger Institutionen auf drei Ebenen von Innovationstätigkeit (Verbundprojekte; Publikationen; Patente). Dazu werden die innovationsbezogenen Datenbanken (s. u.) mittels Record-Linkage-Techniken mit den Daten aus den Institutionendatensätzen Amadeus und Research Explorer verknüpft. Abbildung 2 (S. 6) zeigt strukturiert den Ablauf der Routinen.

Die Ausgangsbasis besteht aus zwei Institutions-Datenbanken (Amadeus und Research Explorer) sowie drei innovationsbezogenen Datensätzen (Förderkatalog; Web of Knowledge; DPMA Deutsche Patente):

1. Einzeldateien Unternehmensdatenbank **Amadeus** Scheiben 2003 – 2014.
   Datenformat: Tab separated values (tsv).
   Variablen: Mark; Company name; BvD ID number; Zip code; City.

2. Tabellen Datenbank **Research Explorer**.
   Datenformat: Excel xlsx.
   Übersicht über deutsche Hochschulen und außeruniversitäre Forschungsinstitute. „Kleine Version“: `20140321_Forschungseinrichtungen_REX.xlsx`; „Große Version“ (mit einzelnen Instituten und Lehrstühlen): `0140507_Forschungseinrichtungen_REX.xlsx`.
   Variablen: Id; Institution; Strasse; Hausnummer; PLZ; Ortsname; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion.

3. Auszug aus dem **BMBF Förderkatalog**.
   Datenformat: Pipe (|) separated values (psv).
   Gefiltert für Projekte mit mindestens einem Akteur aus den sechs RegDemo Fallregionen (vgl. Anhang A.4).
   Inhalt: Projektfördermaßnahmen, Forschungs- und Entwicklungsaufträge.
   Variablen: (Vergleiche auch Übersicht im Anhang 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. Auszug aus **Web of Knowledge** bzw. *Web of Science*.
   Datenformat: Excel xlsx.
   Inhalt: Publikationsdaten (→ Co-Publikationen). Gefiltert für den Zeitraum 2000-2012. Gefiltert nach den sechs RegDemo Fallregionen (s.u.) – nur die Co-Autoren, die sich in den Fallregionen befinden. Alle „externen“ Autoren wurden entfernt.

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

Publikationen: `Gesamtpuplikationen_ROR_innerhalb_ohneSingle_ohneIntern_2.xlsx`. Einzelne Tabellenblätter für Aachen; Dresden; Magdeburg; Kassel; Rostock; Siegen.
Variablen: Pubnr.; ROR Akteure pro Publ.; Akteure pro Pub.; PY; Interne Akteur?; ID; Akteursname.

5. Auszug aus Daten des Deutsche Patent- und Markenamtes (**DPMA**).
   Einzelne Dateien für die sechs RegDemo Raumordnungsregionen. Datenformat: txt/xlsx gemischt.
   Variablen: mainid; patnr_dpma; pa; extern; typassignee; orig_inv; pacity; plz; kgs; ror; bl; pacountry.

3. Aufbereitung der Institutions-Datenbasen

3.1. Aufbereitung Amadeus

In diesem Abschnitt wird die Aufbereitung der Amadeus-Daten beschrieben.

01 Converter.do

Diese 12 Routinen (für eine Lösung mittels Schleife sind die Datensätze in der Struktur zu unterschiedlich) lesen die Amadeus-Daten jahrweise aus den einzelnen tsv-Dateien ein, setzen Variablennamen und Datentypen, erstellen eine Variable für das betreffende Jahr (last_year) und speichern die Dateien im Stata-Format. Die BvDid umfasst 12 Stellen, wobei die ersten beiden Stellen (DE) die Länderkennung sind.

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

Output: `Amadeus_YYYY.dta`

BvDid	Bureau van Dijk ID. Identifikations-ID für Unternehmen
Name	Name des Unternehmens
PLZ	Postleitzahl
Ort	Ort
last_year	Jahrgang des Datensatzes (2003..2014)

02 Append.do

In diesem Schritt werden die einzelnen Jahresdateien mittels append-Funktion zu einem Gesamtdatensatz vereinigt. Dabei werden Duplikate in den Variablen BvDid, Name, PLZ, Ort identifiziert und beseitigt. Der jeweils neueste Datensatz wird behalten.

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

Diese Routine vollzieht das Record-Linkage-Pre-Cleaning (RLPC) auf den Unternehmensnamen (Name) des Amadeus-Datensatzes (zu Details siehe Ehrenfeld 2015c). Dabei werden Sonderzeichen auf ASCII-Zeichen zurückgeführt, Unternehmensformen identifiziert, Klammerausdrücke isoliert und schließlich so ein Ausdruck geschaffen, der ein Vergleichen mit anderen Unternehmensnamen ermöglicht (RLName). Im Zuge dessen werden die Variablennamen an den Datensatz (Präfix BvD - vgl. Anhang A.3) angepasst. Die Angabe des letzten Jahres (last_year) wird auf _lyr verkürzt.

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	Bereinigter und modifizierter Unternehmensname
BvD_legform	Identifizierte Rechtsform des Unternehmens (GmbH, KG, ...)
BvD_temp_name	RLName des letzten Schrittes
BvD_clean_hist	Gibt die durchlaufenen Schrittnummern des RLNamens an
BvD_brackets	Isolierter Klammerausdruck von RLName

04 AGS zuspielen.do

In diesem Schritt wird die Kennziffer des Kreises (AGS5) aus der Postleitzahl bzw. des Ortsnamen generiert und zugespielt. Im Zuge dieses Schrittes werden auch temporäre Dateien des RLPC gelöscht (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	Kreiskennziffer (AGS5)

05 Remove_dup_Name_PLZ_Ort.do

In zwei Stufen werden Duplikate der Daten entfernt. Die erste Stufe bereinigt unterschiedliche Schreibweisen in BvD_Name (zum späteren Vergleichen von Datensätzen wird dann der RLName benutzt – dieser ist zwischen den bereinigten Datensätzen aber identisch). Von den betreffenden Daten wird der neueste Eintrag bzw. der Eintrag mit der größten BvD_ID behalten. Die Duplikate werden anhand von BvD_ID; RLName; BvD_Ort; BvD_AGS5; BvD_legform identifiziert. Die zweite Stufe eliminiert zusätzlich Unterschiede in der BvD_ID bei gleichem Namen, Ort, AGS5 und Gesellschaftsform. Dies ist hier zulässig, da die spätere Zusammenführung anhand des Namens (und anderen Merkmalen, aber nicht der BvD_ID) erfolgt.

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

Dieser Datensatz wird mit den Research Explorer Daten in Schritt 3.2 (`07 Merge\_REX\_Amadeus.do`) vereinigt.

3.2. Aufbereitung Research Explorer

Dieser Abschnitt beschreibt die Aufbereitung der Research Explorer-Daten im Rahmen des Projektes RegDemo. Eine neuere, umfassendere Aufbereitung der Research Explorer-Daten ist in Ehrenfeld (2015b) beschrieben.

01a Converter_20140321.do

Die Daten der „kleinen“ Version des Research Explorer werden aus der xlsx-Datei eingelesen. Das vorhandene Id-Feld wird in eine zum Format von BvD kompatible ID konvertiert (REXid). Die ID umfasst 12 Stellen, wobei die ersten 2-3 Zeichen die Herkunft identifizieren. Hier ist die Herkunftskennung REX. Danach folgen 9 Stellen, die numerisch identisch mit der Nummer aus dem ursprünglichen Id-Feld ist. Die einzelnen Variablen werden bereinigt und formatiert. Das Feld Institution wird um störende Steuerzeichen (CR; LF) bereinigt (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

Die Routine leistet dasselbe wie `01a Converter_20140321.do` für den „großen“ Research Explorer Datensatz. Das Institution Feld wird aber in Institution2 umbenannt, um später beide Datensätze zusammenführen zu können.

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

Da im späteren Verlauf der Zusammenführung Unterschiede in den Schreibweisen von Institutionen aufgefallen sind (z.B. Rheinisch-Westfälische Technische Hochschule Aachen vs. RWTH Aachen), wurde an dieser Stelle eine Tabelle angelegt, die die unterschiedlichen Schreibweisen derselben REX-ID zuordnet. Diese Routine dient dem Einlesen und konvertieren dieser xlsx-Tabelle. Der Funktionsumfang ist identisch mit `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

In diesem Schritt wird die „kleine“ Version des Research Explorer mit den zusätzlich angelegten Schreibweisen vereinigt.

Input: `20140321_Forschungseinrichtungen_REX.dta`
`Forschungseinrichtungen_Add.dta`

Output: `Forschungseinrichtungen_REX.dta`

03 Prepare_Institute_List.do

Für die Zusammenführung nicht benötigte Variablen werden entfernt (Strasse; Hausnummer; Ortsname mit Zusatz; Bundesland; Internetadresse; Fachgebiet; Einrichtungstyp; Sektion). Der Einrichtungstyp wird in eine äquivalente numerische Darstellung (vgl. Anhang A.5) überführt und mit einem der ursprünglichen Schreibweise entsprechenden Label versehen.

Input: `Forschungseinrichtungen_REX.dta`

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

04 AGS zuspielen.do

Den Institutionen werden die Kreiskennziffern (AGS5) und das Bundesland (BuLa) zugespielt. Institution1 wird in Name umbenannt.

Input: `Institutes_Init.dta`

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

05 RLPC_ResearchExplorer.do

Die Namen der Institutionen werden (wie bereits die Namen der Unternehmen) um nicht-ASCII-Sonderzeichen bereinigt und ein Record-Linkage kompatibler Name erzeugt (RLPC). Gesellschaftsformen werden identifiziert; Klammerausdrücke werden abgespalten. Im Zuge dessen werden die Variablennamen an den Datensatz angepasst (Präfix: REX - vgl. Anhang 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

Der Datensatz wird auf die Zusammenführung mit Amadeus vorbereitet. Dazu werden alle Institutions- und Ortsnamen großgeschrieben (Umlaute bleiben erhalten und werden auch großgeschrieben; „ß“ wird nicht verändert – genau wie in Amadeus). Unterschiedliche Schreibweisen von REX_Name werden bis auf die mit der höchsten REX_ID gelöscht. Nicht benötigte Teile aus RLPC werden gelöscht.

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

Der Research Explorer Datensatz wird mit dem Amadeus-Datensatz vereinigt (append). Die Variablennamen werden an den Gesamtdatensatz angepasst (Präfix: ARE für Amadeus/Research Explorer). Duplikate in den Variablen RLName; ARE_Ort; ARE_AGS5 werden gelöscht. Für die Amadeus-Daten wird der Typ (1 = Wirtschaft) nachgetragen. Blocking Variablen für das Bundesland und den Anfangsbuchstaben des Namens werden erstellt (block_name; block_ags). Die resultierende Datei wird im Datenordner von Amadeus gespeichert.

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

Den vereinigten ARE-Daten werden die Informationen zur Raumordungsregion (ROR - siehe Anhang A.4) zugespielt. Die verwendete (externe) Tabelle ist `ROR2011`. Weiter wird eine Zeilennummer (ARE_line) erzeugt, die später für eine fuzzy matching routine benötigt werden könnte (Die Stata-Routine „reclink“ verlangt diese beispielsweise – „reclink“ wurde dann aber doch nicht verwendet). Die csv-Version dieser Tabelle wird später für die Verwendung von „Fuzzy Dupes“ benötigt. Anhang A.2 zeigt die resultierenden Stata-Datentypen und -Formatierungen.

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 einem zweiten Schritt werden alle Datensätze gelöscht, die nicht aus den sechs Fallregionen stammen. Die Zeilennummer (ARE_line) wird neu geschrieben

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

Diese (optionale) Routine teilt den ARE-Datensatz nach Anfangsbuchstaben auf und speichert die Daten blattweise als Excel-Tabelle. Sie ist gut zum händischen Nachrecherchieren von Unternehmen geeignet.

Input: `ARE_2003-2014.dta`

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

Der Datensatz wird im Zuge der Zusammenführung (Schritt 5.1 - `03 Merge ARE IDs.do`) weiterverwendet.

4. Aufbereitung der innovationsbezogenen Datensätze

4.1. Aufbereitung Förderkatalog

Die Organisation des Förderkataloges umfasst für jedes Verbundprojekt eine ID (Ver_nr_vb; später FK_VbNr – als Gruppen ID) sowie eine Vorgangsnummer für jeden beteiligten Akteur (V_nr; später FK_VNr). Dabei ist die Vorgangsnummer auf den Vorgang bezogen – nicht auf den Akteur. Die V_nr stellt also keine Akteurs-ID dar. Im Zuge von Vorarbeiten wurden bereits die Daten aus einer txt-Datei eingelesen; Variablenlabels zugeordnet und Jahresangaben von Beginn und Ende (year_begin; year_end) erstellt.

Weiter wurden ältere Angaben der Gemeindekennziffern aktualisiert (Gebietsreform 2011) und Daten für die Raumordnungsregion zugespielt. Anschließend wurden die Daten gefiltert: a) Nur Projekte mit Nähe zu FuE; b) Nur Daten mit Angaben zur ROR; c) Jahre 1991-2010; d) Nur Verbundprojekte. Schließlich wurden nur jene Projekte behalten, bei denen

mindestens ein Akteur aus einer der sechs RegDemo-Fallregionen stammt. In diesem Fall wurde aber der komplette Datensatz erhalten (`Foeka_Fallregionen_WE.dta`)

01 Foeka_Fallregionen_Institutionen_Prepare_RL.do

Diese Routine bereitet die Förderkatalogdaten auf die RLPC Routine vor. Dazu werden die Variablennamen an den Datensatz angepasst (Präfix: FK für **F**örder-**K**atalog); Großschreibung wird verwendet für den Namen der ausführenden Stelle (Name_St); den Ort der ausführenden Stelle (Ort_St); sowie den Namen und Ort des Zuwendungsempfängers (Name_ZE; Ort_ZE). Für die spätere Zusammenführung ist jedoch die ausführende Stelle entscheidend.

Der Typ (FK_Typ) der Organisation wird in ein numerisches Format konvertiert und gelabelt (vgl. Anhang A.5). Weiter werden sequenziell folgende Filter angewendet: a) Nur Akteure aus den Fallregionen (FK_FR - an dieser Stelle werden alle externen Akteure verworfen, die zusammen mit internen Akteuren kooperiert haben); b) Nur die Jahre von 2000-2012. Um den Namen der obersten Institutsebene zu erhalten wird der Name der Stelle (FK_Name_St) aufgespalten und - wenn diese Bezeichnung zu kurz ist - um einen weiteren Teil des Stellennamens ersetzt.

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; 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

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 einem zweiten Schritt werden die Akteure auf Stellenebene isoliert. Hier ist z.B. noch die Fakultät unterscheidungsrelevant für den Akteur.

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

Im dritten Schritt wird nur noch die oberste Ebene der Institution beibehalten. In diesem Schritt wird auch eine Akteurs-ID für diese Ebene vergeben (FKA_ID). Die IDs sind 12stellig; fortlaufend nummeriert (Sortierung FK_Name; FK_Ort; FK_AGS5) und tragen den Präfix 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

Diese Routine führt die RLPC-Routine auf die FK-Institutions-Daten aus.

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 diesem Schritt werden den Vorhaben die IDs der beteiligten Akteure zugeordnet. Die Vorhabensdaten stammen aus `Foeka_Fallregionen_Vorhaben.dta` (vgl. Schritt 01). Die IDs stammen aus `Foeka_Fallregionen_Institutionen_RLPC` (vgl. Schritt 02). Anhang A.2 zeigt die resultierenden Stata-Datentypen und -Formatierungen als Beispiel für den strukturellen Ausbau der Datensätze.

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 einem zweiten Schritt werden die Daten nach Fallregionen <FR> aufgeteilt.

Output: `Foeka_FKA_<FR>`

Dieser Datensatz wird in Schritt 5.1 - `01 Zusammenführen.do` mit den Bibliometrie-Daten und den Patent-Daten zusammengeführt.

4.2. Aufbereitung Bibliometriedaten

Für jede Publikation gibt es eine einheitliche Publikationsnummer (Pubnr., später WK_PubNr). Diese wird als Gruppen-ID verwendet.

01 ID_ROR_7_RLPC.do

Die Routine erstellt Schlüssel für die in der Akteursliste (ID_ROR) verwendeten numerischen Codes (ID_ROR_7_Cities; ID_ROR_7_Legforms) und ordnet den Codes die Einträge im Klartext zu. Sie liest die Akteursdaten aus der tsv-Datei ein; wandelt die bereits vergebenen Akteurs-IDs in das einheitliche Format (WKA_ID; Präfix: WKA für **W**eb-of-**K**nowledge-**A**kteur) und filtert Duplikate der Akteurs-IDs. Weiter werden die Einträge für den Organisationstyp (WK_Typ) auf das in den anderen Datensätzen verwendete Format gebracht. Name und Ort werden - wie bei den anderen Datensätzen - großgeschrieben. Schließlich wird der Akteursname (WK_Name) der RLPC-Routine unterzogen ($\rightarrow$ 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

Die Publikationsdaten der einzelnen Fallregionen <FR> werden eingelesen. Die bereits vorhandene ID wird in das einheitliche Format überführt (WKA_ID). Die Daten zu den Akteuren aus `ID_ROR_7_RLPC.dta` (aus Schritt 01) werden zugespielt.

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

Dieser Datensatz wird in Schritt 5.1 - `01 Zusammenführen.do` mit den Förderkatalog-Daten und den Patent-Daten zusammengeführt.

4.3. Aufbereitung Patentdaten

Jedes Patent trägt eine Patentnummer (patnr_dpma; DP_DPMA). Weiter beinhaltet der Datensatz eine eigene ID (mainid; DP_Mainid), deren Abdeckung in diesen Datensätzen besser als ist die Patentnummer und deshalb als Gruppen-ID verwendet wird. Die Datensätze sind für die Fallregionen untergliedert und beinhalten den Namen des Patentanmelders (pa); den Typ des Anmelders (1 = Wirtschaft; 2 = Hochschule; 3 = außeruniversitäre Forschungseinrichtung; 4 = sonstige - vgl. Anhang A.5); das Originalfeld des DPMA-Datensatzes für den Anmelder (orig_inv; DP_Orig) sowie Ortsnamen, Kreiskennziffer/AGS5, PLZ, ROR, Bundesland (bl; DP_BuLa) und Land (pacountry; DP_Ctry) für den Anmelder. Die Variable bl wurde aus der ROR generiert.

Weiter gibt es eine Anzeige (extern; DP_Extern), warum der Datensatz aufgenommen wurde. Sie gibt an, ob der Anmelder aus der ROR stammt (extern = nein/0), oder ob mindestens ein Erfinder aus ROR stammt, aber nicht der Anmelder (extern = ja/1). Die Abdeckung der Ortsnamen ist aber oft ein Problem.

Ein Problem in den Originaldaten ist die räumliche Zuordnung der Innovationsleistung. So wurden beispielsweise sämtliche Patente von Audi in Ingolstadt angemeldet, unabhängig davon, wo die Erfindung gemacht wurde. Oftmals existiert aber eine Zweigstelle dort, wo

die Erfindung gemacht wurde. In diesen Fällen wurde der Name und die ROR (und abhängig davon das Bundesland - aber nicht die anderen Daten!) so verändert, dass der Name nun die regionalspezifische Stelle enthält (z.B. Audi AG Kassel).

01 Einlesen.do

Die Patentdaten werden aus den tsv-Dateien eingelesen (Die Originaldaten waren xlsx-Dateien – diese wurden in tsv-Dateien überführt). Die Labels für den Akteurstyp (DP_Typ) werden vereinheitlicht und Daten für die Raumordnungsregionen zugespielt (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

Akteursname und Ortsname werden in Großbuchstaben konvertiert. Anschließend werden nur die Einträge behalten, an denen Akteure aus den Fallregionen beteiligt sind. Weiter werden alle Einzeleinträge gelöscht (Ziel sind ja Kooperationsbeziehungen).

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 einem zweiten Schritt werden die Patenteinträge aufgelöst und nur die um Duplikate (DP_Name; DP_Ort; DP_AGS5) bereinigten Akteure behalten.

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

Im dritten Schritt werden die einzelnen Akteurslisten der Raumordnungsregionen zu einer zentralen Liste vereinigt. Den Akteuren wird eine ID nach vereinheitlichtem Format (Präfix DPA für Deutsches Patent Akteur) zugeordnet (Sortierreihenfolge: 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

Für die Akteursnamen wird ein Record-Linkage-fähiger Name erzeugt (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

Den Patenten aus den Fallregionen aus Schritt 02 (`DP_ROR_Filter_<FR>.dta`) werden die DPA-IDs aus Schritt 03 (`DP_Akteure_RLPC.dta`) zugespielt.

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

Dieser Datensatz wird in Schritt 5.1 - `01 Zusammenführen.do` mit den Förderkatalog-Daten und den Bibliometrie-Daten zusammengeführt.

5. Aufbau der Innovationsnetzwerke

5.1. Zusammenführen der Datensätze

Nachdem die Einzelkomponenten aufbereitet wurden, werden diese nun zusammengeführt.

01 Zusammenführen.do

In dieser Stufe werden die Akteure der drei innovationsbezogenen Datensätze (Förderkatalog; Bibliometriedaten und Patentdaten) zusammengeführt (`Foeka_Fallregionen_Institutionen_RLPC` aus Schritt 4.1; `ID_ROR_7_RLPC` aus Schritt 4.2 und `DP_Akteure_RLPC` aus Schritt 4.3). Amadeus- und Research Explorer Daten wurden bereits in Schritt 3.2 - `07 Merge_REX_Amadeus.do` miteinander vereinigt. Die einzelnen ID-Spalten werden zur einheitlichen Spalte ID vereinigt. Hier finden auch einige Modifikationen des RLNamens statt um bestimmte Schreibweisen mit Zusätzen zu erfassen.

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

Dieser Schritt umfasst zwei Stufen, welche die externe (kommerzielle) Software „Fuzzy Dupes“ verwenden. Die Einstellungen dieses Programms befinden sich im Anhang A.6 und A.7. In der ersten Stufe werden der vereinigten Akteursliste aus Schritt 01 die Zeilennummern der korrespondierenden Einträge aus dem kombinierten ARE-Datensatz zugeordnet.

Die im Zuge der Zuordnung gebildete FuzzyMatchID bezieht sich dabei auf die Zeilennummer(!) des ARE-Datensatzes; _Line# ist die Zeilennummer des geladenen Akteurs-Datensatzes; Matching gibt einen (numerischen) Matching-Score [0;1] an und steht somit für die Güte der Passung zwischen Akteurs-Datensatz und ARE-Datensatz .

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 der zweiten Stufe werden Duplikate von Akteuren identifiziert und mit einer Fuzzy Dupes eigenen Gruppen-ID versehen. FuzzyDupesID ist die neu gebildete Gruppen ID. Akteure, welche dieselbe ID tragen, sind vermutlich identisch; Matching ist wieder der Matching-Score; mit Delete werden zu löschende Einträge gekennzeichnet; dieser Parameter wird hier aber nicht verwendet.

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. Vereinheitlichen der IDs

03 Merge ARE IDs.do

In der ersten Stufe wird die vereinigte Akteursliste mit dem kombinierten ARE-Datensatz vereinigt („left join“). Als Schlüssel dient hier die ARE-Zeilennummer. In der zweiten Stufe werden Duplikate von Akteuren identifiziert und mit einer Gruppen-ID versehen. Schließlich wird eine vereinheitlichte ID vergeben (U_ID für „**U**nified **ID**“).

Im Zuge der Zusammenführung und Duplikateverarbeitung werden einige Variablen umbenannt bzw. neu erzeugt: Die korrespondierende ARE-Zeilennummer wird von FuzzyMatchID in mline umbenannt (für matching line – wird später gelöscht); hiermit wird u.a. die ARE_ID gemerged; ms bezeichnet den Matching-Score von Akteursliste und ARE-Datensatz.

Der Matching-Score für Akteursduplikate ist ds; cFD zählt die Einträge in einer Fuzzy Dupes Gruppe. Aus der Gruppierung von Fuzzy Dupes IDs wird mit FDG_ID eine neue Gruppen-ID erstellt; cml zählt (in Analogie zu cFD) die Anzahl der Einträge, die gleichzeitig die selbe FDG_ID und die selbe matching line (zu ARE) haben. Der Vergleich von cFD und cml dient der Aufdeckung von Diskrepanzen zwischen Gruppenbildung durch Duplikate und dem Match zu ARE.

Die vereinheitlichte ID (U_ID) wird hierarchisch vergeben: Falls für den Eintrag eine ARE-ID existiert, so wird diese übernommen. Existiert keine ARE, so wird die Gruppen-ID (FDA_ID) übernommen. Sollte auch diese nicht vorhanden sein, so wird die ursprüngliche ID aus dem Akteursdatensatz übernommen (FKA/WKA/DPA); cUID zählt die Einträge pro Gruppe mit derselben UID. Es kann vorkommen, dass diese Gruppengrößen höher ausfallen als cFD und cml, da die Zuordnung von ARE und Identifizierung von Duplikaten unabhängig voneinander erfolgt, letztendlich aber zur selben UID führt.

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

Für ein vereinfachtes Zuspielen zu den einzelnen Innovationsdatensätzen wird eine Tabelle erstellt, die nur die Original-ID (ID) enthält, sowie die vereinheitlichte ID (U_ID).

Output: `Akteure_FD_ARE_UID.dta`
ID; U_ID

04 Vorhaben UIDs.do

In diesem Schritt werden die U_IDs auf die Innovationsdatensätze in ihren Verzeichnissen zurückgespielt. Die ursprünglichen Gruppenbezeichnungen (FK_VbNr; WK_PubNr; DP_Mainid) werden in die einheitliche Variable Group überführt. Die Akteurs-IDs werden als Node geführt; ihre Namen als Node_Name. Node_nr gibt die laufende Nummer des Knotens innerhalb ihrer Gruppe an.

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. Aufbau der Netzwerkstrukturen

05 Networks.do

Diese Routine erstellt jeweils vollverknüpfte, ungerichtete Netzwerke aus den drei Innovationsdatensätzen. Die Kanten bestehen dabei aus jeweils zwei gegenüberstehenden Knoten (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

In der zweiten Stufe des Schrittes werden die Frequenzen/Häufigkeiten der resultierenden Kanten ausgezählt.

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

06 Comparison Table.do

In dieser Stufe werden die Häufigkeitsverteilungen der drei Innovationsdatensätze für jede Raumordnungsregion zusammengefasst. Sie dienen dem Vergleich der Innovationstätigkeit für jede Kante der zuvor erstellen Netzwerke.

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 einem zweiten Schritt werden die Tabellen über alle Raumordnungsregionen zu einer Tabelle zusammengefasst.

Output: `Comparison_Total.dta`

Literatur

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

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

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

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

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

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

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

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

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

A. Anhang

A.1. Variablen Förderkatalog

Name	Bedeutung
V_nr	Vorhaben-Nummer (eindeutiges Schlüsselfeld)
FKZ	Förderkennzeichen
V_disref	Referat des jeweiligen Ressorts
V_ressort	Ressort
V_thema	Thema des Vorhabens
V_foeart	Förderart
V_pt	Projektträger
V_proref	Zuständige Stelle des Projektträgers
Fi_von	Laufzeit-Beginn
Fi_ende	Laufzeit-Ende
Fi_sumbew	Gesamtbewilligungssumme des Vorhabens
lpsys	Leistungsplan-Nummer
lptext	Leistungsplan-Beschreibung
Name_ze	Zuwendungsempfänger
Ort_Ze	Ort des Zuwendungsempfängers
Bundesland_ze	Bundesland des Zuwendungsempfängers
Gemkz_ZE	Gemeindekennziffer des Ortes des Zuwendungsempfängers
Gembz_Ze	Gemeindename auf der Gemeindekennzifferebene zum Zuwendungsempfänger
Land_ZE	Land des Zuwendungsempfängers (falls nicht in Deutschland)
Name_St	Ausführende Stelle
Ort_St	Ort der ausführenden Stelle
Bundesland_St	Bundesland der ausführenden Stelle
Gemkz_St	Gemeindekennziffer des Ortes der ausführenden Stelle
Gembz_St	Gemeindename auf der Gemeindekennzifferebene zur ausf. Stelle
Land_St	Land der ausführenden Stelle (falls nicht in Deutschland)
Hist_prof	Förderprofil
Hist_prof_Klar	Förderprofil-Beschreibung
Wahlkreis	Wahlkreis-Nummer (aktuelle Legislaturperiode)
Wzweig	Wirtschaftszweig-Zuordnung
Ver_nr_vb	Verbund-Kennzeichen (Schlüsselfeld zur Datei IWH_Verbund)
Ste_key_ZE	Stellensystematik des Zuwendungsempfängers (Schlüsselfeld zu Datei IWH_Stesys)
Ste_key_St	Stellensystematik der ausführenden Stelle (Schlüsselfeld zu Datei IWH_Stesys)
V_Stesys	Stellensystematik auf Vorhabenebene (Schlüsselfeld zu Datei IWH_Stesys)

A.2. Datentypen und Darstellung – Prinzipieller Aufbau

Name	Typ	Formatierung	Kommentar
Beispiel Akteure: [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	
Beispiel Innovatios-Vorhaben: [Foeka_Fallregionen_Vorhaben_FKA.dta]			
FK_VbNr	long	%10.0g	Gruppen ID
FK_VNr	long	%9.0g	
FKA_ID	str12	%12s	Akteurs 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. Herkunft Daten in IDs

Format: 12 Stellen

Präfix	Bedeutung
DE	Amadeus Deutschland
REX	Research Explorer
FDA	Durch Fuzzy Dupes gruppierte Akteure
WE	Zur Gruppierung händisch zugeordnet
FKA	Förderkatalog Akteur
WKA	Web of Knowledge Akteur
DPA	Deutsches Patentamt Akteur

A.4. Fallregionen RegDemo

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

A.5. Codierung Akteurstypen

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

A.6. Fuzzy Dupes Schritt 1: Zusammenführen mit ARE-Datensatz

* 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 Schritt 2: Identifikation von Duplikaten –> Gruppen-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