Disambiguation of author addresses in bibliometric databases – technical report –

Christine Rimmert, Holger Schwechheimer,
Matthias Winterhager

This report has been generated from an internal documentation about the institutional disambiguation developed and processed by the Bibliometric Group of Bielefeld University (part of the Institute for Interdisciplinary Studies of Science – I^2SoS)[1] since 2008 in the context of the German Competence Centre for Bibliometrics[2]. Results of this disambiguation process are an integral part of the quality assured and standardized databases built up from Web of Science and Scopus raw data. Although the application is restricted to Web of Science and Scopus databases in this context, the procedure may also be applied on other data sources.

In the following 'institutional disambiguation' is defined as the assignment of author addresses (recorded in bibliometric databases) to real existing research institutions. Aspects to be taken into account here are not restricted to variance in author addresses due to e.g. name changes, spelling variants, abbreviations or different languages but also include structural changes of institutions over time such as fusions, outsourcing, incorporations or splits as well as data quality aspects in the data sources (e.g. errors or incomplete addresses).

1 Introduction

The disambiguation of institutional addresses is a process that consists of several phases[3]. The core of the process is the allocation of data sets of addresses to the relevant institutions (research institutions and their sub-units) by means of an inventory of regular expressions (in the following called 'patterns') that

[1] http://www.uni-bielefeld.de/i2sos/bibliometrie/ (accessed 14.07.2017)

[2] http://www.bibliometrie.info/ (accessed 14.07.2017)

[3] The basics of the project are described in Winterhager, M., Schwechheimer, H., & Rimmert, C. (2014). Institutionenkodierung als Grundlage für bibliometrische Indikatoren. Bibliometrie - Praxis und Forschung, 3(14), 1–22.

was created with great manual effort. This inventory currently contains 51,600 patterns[4] for Germany and is constantly being maintained and extended.

The cornerstone of the procedure is the application of patterns. This is preceded by preparatory steps and followed by the allocation to the respective main institution taking into account information about hierarchical relationships and structural changes of the institutional landscape over time.
In order to be able to take different requirements (depending on the context of application) into consideration with regard to structural changes, the procedure is designed in a way that different modes of allocation can be chosen. In accordance with the needs articulated so far, two variations of the allocation process are developed:

- Mode A ('current perspective'): Allocation according to the current institutional situation
- Mode S ('synchronic allocation'): Taking into account the historical situation during the year of the respective publication[5]

Other variations for certain contexts of application are possible and can be provided if needed, and if corresponding conditions are formulated.
Moreover, an aggregation of the allocation is conducted and made accessible on the level of the large sectors of the German system of science (universities, Fraunhofer-Gesellschaft, Helmholtz-Association, Max-Planck Society, Leibniz Association etc.) since it plays an important role in the context of many projects. Changes over time and multiple allocations have to be taken into account here as well. The procedural steps in the disambiguation of institutional addresses can be roughly classified into three blocs:

- Preparation of address data,
- Application of pattern recognition and
- Aggregation and result processing.

In the following, we will first present data that are drawn from external sources and which are necessary for the procedure (called 'basic data' hereafter). Subsequently, the relevant steps of the procedure within these three phases are explained in more detail, prerequisites and problematic cases are pointed out, and examples are presented.

[4] July 2017

[5] For these two modes, the results of the allocation are also fed into the databases of the national Competence Centre for Bibliometrics. For details, see Winterhager et al., ibid., p. 9.

2 Basic data

In order to allocate addresses from bibliometric databases to existing research institutions, more information on the objects of classification (the research institutions) is necessary. This information is not available in the bibliometric databases but has to be obtained from external sources – a relational database is used for recording and storage. In general, this concerns

- Characteristics of research institutions (as, e.g., name, date of founding and possible shut down, URL, postal address) and sectors (such as name, URL, possible further classification into sub-sectors)

- Relationships between research institutions, on the one hand, and between sectors and research institutions, on the other hand (sector allocation)

whereas relationships between research institutions can be of different forms, for example, hierarchical relationships, predecessor-successor relationships, relationships between affiliated institutes and universities, between teaching hospitals and universities, networks, unions or umbrella organizations.

The modelling of the basic data in the relational database is designed in a way so that it can include all types of relationships between units/entities, even though a complete collection cannot be achieved and not every relationship has an influence on the actual procedure, i.e. the disambiguation of institutional addresses. Since the application of the procedure is conceivable in very different contexts, and thus under very different prerequisites, the goal is to provide a strong degree of flexibility.

All data on characteristics and relationships are labelled with dates in order to be able to trace changes over time (for example, a structural change such as the outsourcing of a former sub-unit or the name change of a research institution).

An additional group of basic data is represented by transformation and allocation rules (referred to here as ‘transformation’) for addresses. These are also necessary for the procedure and will be described in more detail in the following.

Figure 1 shows a simplified entity-relationship-model of the basic data.

The central type of entity is the unit. Here, all hierarchical levels of research institutions can be recorded: a working group is a unit, as is a faculty or a university. Units are identified with a start and end date, which correspond to the foundation, resp. shut down. An internal ID is assigned and – if existing – a URL is identified.

Universities, universities of applied science, the Max-Planck Society (MPG), the Fraunhofer-Gesellschaft (FhG), the Leibniz Association (WGL), the Helmholtz-Association (HGF) etc. as well as others such as hospitals, enterprises, Federal and State Government R&D institutions, are identified as sectors.

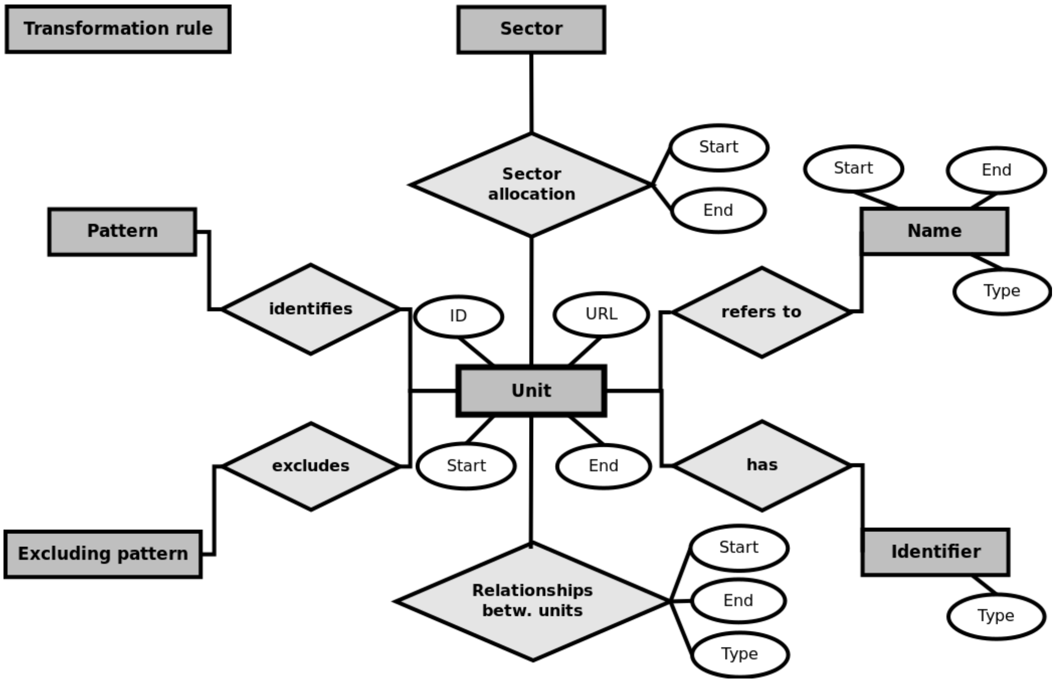


Figure 1: Entity relationship model of basic data.

The allocation of units to sectors is depicted as a relationship and multiple allocations (for sector-hybrid institutions) are possible. The relationships are labeled with start and end dates in order to be able to trace sectorial changes. Relationships, however, not only exist between units and sectors but also between different units. Different kinds of relationships can be identified via the type (not only hierarchies, but also relationships between affiliated institutes and universities, clinics and universities in case of teaching hospitals etc.). Relationships between units are also labeled with start and end dates (resp. in the case of predecessor-successor relationships with transitional date – this differentiation is not explicated in the ER-diagram due to simplification). An outsourcing is thus depicted, for example, if the hierarchical relationship to the superior unit receives an end date.
Names of units are also identified by a start and end date (in this way name changes can be identified) and a type (therefore, different variations of a name can be recorded – e.g. German and English names, abbreviations, alternative descriptions etc.).
Moreover, further identifiers can be recorded in order to create concordances to other databases. The patterns and excluding patterns referring to units include concrete allocation rules for addresses and will be described in more detail below (as will the transformation rules for addresses).

3 Preparation of the address data

Preparatory steps include the selection, extraction and transformation of the relevant data sets from the respective raw data material (e.g. from Web of Science or Scopus). Figure 2 uses the example of an unproblematic address to demonstrate an overview of the steps in the preparation of address data. On the left side the necessary prerequisites are listed, inputs for the steps are depicted in the middle, and the address example is shown on the right side.

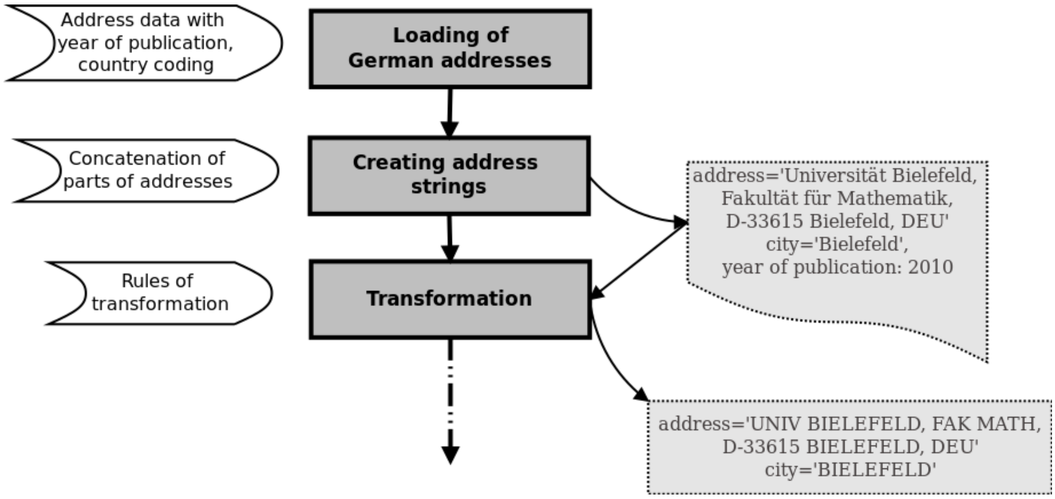

Address data with year of publication, country coding
Loading of German addresses
Concatenation of parts of addresses
Creating address strings
address='Universität Bielefeld, Fakultät für Mathematik, D-33615 Bielefeld, DEU'
city='Bielefeld',
year of publication: 2010
Rules of transformation
Transformation
address='UNIV BIELEFELD, FAK MATH, D-33615 BIELEFELD, DEU'
city='BIELEFELD'

Figure 2: Preparation of address data.

3.1 Selection and loading of the address data sets

The selection of the pertinent address data sets from the raw data material is initially steered via the country data available in the respective source. Since this data field is not always sufficiently standardized, a country coding is used that was developed solely for this purpose and which takes into account all variations (for Germany, for example, *FED REP GER*, *WEST GERMANY*, *DEUTSCH DE*, *Alemania*, *BERLIN*, *Deustchland*, *GER DEM REP*, ...). The procedure is so far primarily designed for the allocation of 'German' publications, i.e. publications are selected where at least one author has provided an institutional affiliation in Germany. For the subsequent steps in the procedure, identifiers for publication, year of publication, identifiers for address, address string and city string (optional) are loaded and integrated into the further processing.
In order to enable, resp. optimize, allocation via pattern recognition, address data sets are subject to a prior step where the separately available parts of addresses (if necessary) are assembled and transformed.

3.2 Assembling parts of addresses

For the subsequent process of pattern recognition, the addresses are necessary as input in form of a connected string and optionally another string with information on the city[6]. A connected address-string is not always available in the raw data (in the case of Scopus, for example, only parts of addresses are available in separate data fields, and a string that contains the entire address, such as in the Web of Science raw data in XML format in the data field *FULL_ADDRESS*, is missing). In these cases, the separate parts are integrated into one string (separated by commas), if possible in the sequence that is standard in the Web of Science: Name of organization, name of the sub-unit(s), street name/post box, city, country.

3.3 Transformation

The number of the applied patterns is a significant parameter for the efficiency of the process of pattern recognition. The development and maintenance of text patterns for the automatic recognition of addresses entails significant (manual) effort. The growing inventory of patterns is accompanied by an increasing workload with respect to quality control. Furthermore, an unnecessary large inventory of patterns has a negative effect on the duration of the processing. Therefore, the allocation of addresses to units via the patterns is preceded by a procedural step where variations of sub-strings are taken into account, resp. are standardized, which can have an effect on the allocation of addresses for several target units (in contrast to the subsequent application of pattern recognition where it is solely about one target unit in each case). In this context, certain sub-strings are replaced by a standardized form, for example:

Universität, *University of*, *Universitat*, *Universidad*, ... → *UNIV*.

This transformation has effects on addresses of many units for which otherwise individual text patterns would have to be developed. In addition, this procedural step of transformation can also serve to correct obvious misspellings (such as *Universitt*) in general from the beginning (regardless of the target units of the subsequent allocation).

The rules for the transformation are fed by different sources; they are oriented towards the pre-standardization[7] of the Web of Science but not restricted to them. Aside from rules for certain endings, for example

[6] Such strings for a city can be found e.g. in the data fields *CITY_GROUP* in Scopus and *CITY* in the Web of Science.

[7] cf. the sections 'Address Abbreviations' and 'Corporate and Institution Abbreviations' in the documentation 'Web of Science Core Collection Help'. URL: http://images.webofknowledge.com/WOKRS517B4/help/WOS/index.html (accessed 14.07.2017)

$$OLOGIE \text{ and } OLOGY \rightarrow OLOG,$$

individually and manually identified replacements (e.g. for misspellings) as well as the acceptance of replacements resulting from the use of the statistical translation tool MOSES [8] (after manual check) are applied. The transformation is not limited to replacing individual words. Longer sections are transformed, parts are permuted, stop-words deleted, special characters are transformed as well etc.
The transformation is carried out on the address string as well as on the city string. Table 1 shows the results of the transformation of the address string (without transformation of the city string) for several examples.

Before Transformation	After Transformation
University of Regensburg, Institute of Physical and Theoretical Chemistry, D-93040, DEU	UNIV REGENSBURG, INST PHYS & THEORET CHEM, D-93040, DEU
Max Planck Institute for Infection Biology, Department of Immunology, Schumannstr. 21/22, 10117 Berlin, DEU	MAX PLANCK INST INFECT BIOL, DEPT IMMUNOL, SCHUMANNSTR 21 22, 10117 BERLIN, DEU
Ludwig-Maximilians-Universität, Department of Earth and Environmental Sciences, Geophysics Section, München, DEU	LUDWIG MAXIMILIANS UNIV, DEPT EARTH & ENVIRONM SCI, GEOPHYS SECT, MUNCHEN, DEU
Technische Universitt Mnchen, Walter Schottky Institut, Am Coulombwall 3, 85748 Garching, DEU	TECH UNIV MUNCHEN, WALTER SCHOTTKY INST, AM COULOMBWALL 3, 85748 GARCHING, DEU

Table 1: Examples for the transformation of address strings.

[8] cf. MOSES - Statistical Machine Translation System.
URL: http://www.statmt.org/moses/ (accessed 14.07.2017)

4 Application of pattern recognition

Figure 3 shows the sequence of the steps that follow the preparation of address data. These entail the identification of text patterns and form the core of the entire procedure (the disambiguation of institutional addresses). Analogous to figure 2, the left side depicts the necessary inputs for the steps (shown in the middle), while the right side presents the effects on the example.

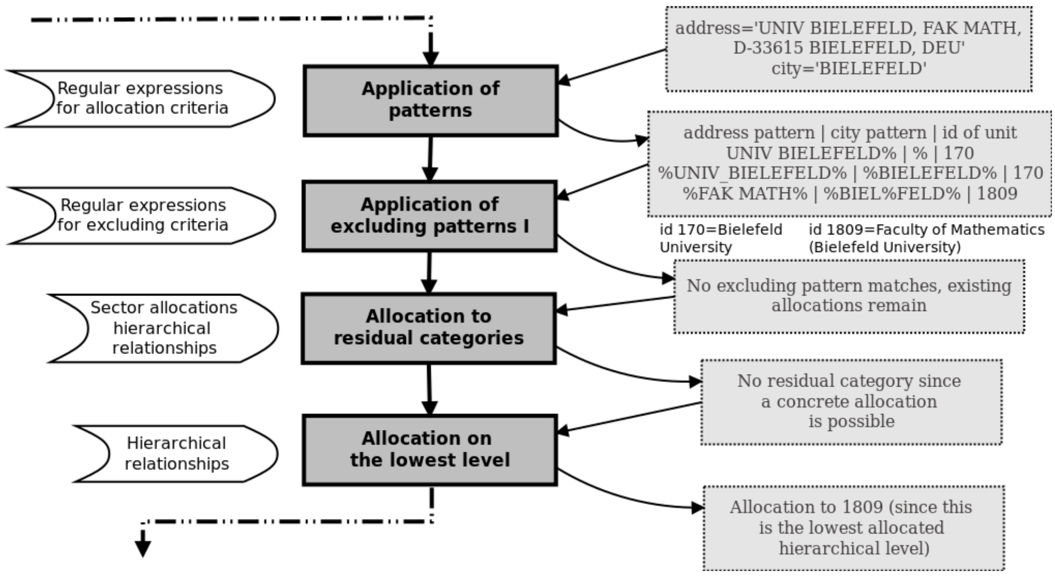

address='UNIV BIELEFELD, FAK MATH, D-33615 BIELEFELD, DEU' city='BIELEFELD'
Regular expressions for allocation criteria
Application of patterns
address pattern | city pattern | id of unit
UNIV BIELEFELD% | % | 170
%UNIV_BIELEFELD% | %BIELEFELD% | 170
%FAK MATH% | %BIEL%FELD% | 1809
id 170=Bielefeld University
id 1809=Faculty of Mathematics (Bielefeld University)
Regular expressions for excluding criteria
Application of excluding patterns I
No excluding pattern matches, existing allocations remain
Sector allocations hierarchical relationships
Allocation to residual categories
No residual category since a concrete allocation is possible
Hierarchical relationships
Allocation on the lowest level
Allocation to 1809 (since this is the lowest allocated hierarchical level)

Figure 3: Application of pattern regognition.

While the objective of the procedure is to allocate publications (on the basis of the affiliations provided by the authors) to the respective main institution[9] (university, Max-Planck Institute etc.), the identification of patterns and the allocation of addresses can occur on every level in the hierarchy of an institution. The inventory of patterns includes both those patterns that allow a precise allocation to a certain working group, an institute or a faculty/department, as well as those that merely allow a general allocation to a university as a whole or on the sector level (for these allocations special classification targets called 'residual categories' have been created).

A complete identification of every sub-unit of all main institutions of the German system of science is not required (and would hardly be possible with a reasonable amount of work). The procedure and the arrangement of the basic tables in the database scheme, however, are designed in a way so that information on the internal differentiation of main institutions (where available) can be included and taken into account at any time. In certain cases, infor-

[9] On the definition of the main institution, cf. Winterhager et al., ibid., p. 7.

mation about sub-units of a main institution is essential in order to carry out allocations to the main institution. This is especially true for the handling of addresses where only a sub-unit, not the corresponding main institution, is explicitly named. One example is the following address:

Center of Excellence in Cognitive Interaction Technology (CITEC), Bielefeld, DEU.

The prerequisite for the correct allocation to the correct main institution is the information that CITEC is a sub-unit of Bielefeld University.

4.1 Principles of creating patterns

The patterns used for text recognition in the procedure consist of two regular expressions per pattern[10], one of which is applied to the address string (hereafter 'address pattern') and the other to the city string (hereafter 'city pattern'). To avoid mistakes in the manual production and maintenance of the pattern inventory by several persons, only very simple regular expressions were allowed in the development phase. The following wildcards were exclusively used:

- precisely one character ('_')
- any number of characters ('%')

Since there is now an extensive inventory of patterns and fewer people are involved in the identification of new patterns, gradually more complex regular expressions are developed from the existing regular expressions. These enable a summary of several simple regular expressions into a single regular expression, so that the inventory can be reduced significantly while keeping its level of functionality. In this context, the opportunities provided by the more complex regular expressions should be taken advantage of, but should also remain readable (see principles of creating patterns).
The application of the patterns is done in a relational database system via SQL commands with 'WHERE'-clause and 'LIKE'-operator (where '_' and '%' become effective as wildcards) resp. REGEXP_LIKE for more complex terms. In the case of an address without any available city string, only patterns are used whose city pattern ='%'. Table 2 shows some examples of address and city patterns for the RWTH Aachen.

[10] Cf. Kevin Loney (2009). Oracle Database 11g, Die umfassende Referenz, p. 147-161

Address pattern	City pattern
AACHEN UNIV%	%AACHEN%
%UNIV TECH AIX LA CHAPELLE%	%
TH_AACHEN%	%
RHEIN WESTFAL TECH HSCH%	%AACHEN%
WESTFAL TECH HSCH,%	%AACHEN%
RTW AACHEN UNIV,%	%AACHEN%
AACHEN TECH HSCH,%	%AACHEN%
RHINE WESTFALIA TECH UNIV%	%AACHEN%
TECH HOCHSCHULE AACHEN,%	%AACHEN%
RWTH AACHEN%	%

Table 2: Examples of patterns for RWTH Aachen.

Certain criteria apply to the creation of patterns: the patterns should be constructed

- as simple as possible (avoiding mistakes, overview),
- as general as possible (recall) and
- as specialized as necessary (precision).

These criteria, however, can only serve as guidelines. Whether they are fulfilled eventually cannot be evaluated since the amount of hits for a pattern depends on the actual raw data of the addresses that is available and can therefore, in the case of the Web of Science, change on a weekly basis. A pattern which only achieves correct hits in one data inventory can trigger errors in another data inventory (because this inventory possibly contains other addresses).

Therefore, the continuous control and maintenance of the pattern inventory is crucial. The evaluation of the coding can provide indications regarding necessary changes or additions in the pattern inventory. Thus, the systematic analysis of the remaining quantity of addresses that have not yet been allocated can contribute to the identification of additional candidates for which new patterns are necessary and useful. Tests of the quantity of completed allocations with regard to errors can provide information on necessary changes for existing patterns. Since the disambiguation of institutional addresses is frequently applied in bibliometrics for obtaining data on publications (and dependent indicators) as well as on citations (and dependent indicators), address variations that often appear (i.e. are used in many publications) or belong to highly cited publications have priority in the design of patterns. This is supposed to achieve a broad coverage of the ‘important’ addresses.

4.2 Excluding patterns

In some cases it is helpful or even necessary to define excluding patterns (especially due to the use of simple regular expressions and the need of keeping the number of patterns limited and in the same time achieve the necessary demarcations between units with similar names). These also consist of an address pattern and a city pattern and are used as follows: if a pattern (A) fits a certain allocation target and an excluding pattern (B) fits the same target unit at this address, then the allocation will not be carried out via the pattern (A) (resp. it is deleted). This often occurs for teaching hospitals. As these hospitals are not part of the university, an automatic allocation to the university is not justified[11]. An excluding pattern that contains the string which describes the teaching hospital can prevent the corresponding allocation. The following example helps to explain this:

- Address string: *UNIV MUNSTER, AKAD LEHRKRANKENHAUS, CLEMENSHOSPITAL, DUESBERGWEG 124, D-48153 MUNSTER, DEU*,
- City string: *MUNSTER*

A pattern designed for the University of Münster (Westfälische Wilhelms-Universität Universität Münster) (ID=100) matches this address:

- Address pattern: *UNIV MUNSTER%*
- City pattern: *MUNSTER*

and leads to the corresponding allocation. At the same time, an excluding pattern fits the ID 100 (Universität Münster):

- Address pattern: *%LEHRKRANKENHAUS%*
- City pattern: *%*

and thus overrides the allocation to the University of Münster.

Another example of where excluding patterns are applied is represented by addresses for which the pattern *UNIV FRANKFURT%* fits. Here, an allocation to Goethe University (Frankfurt a.M.) should not occur if the city string fits the excluding pattern *FRANKFURT ODER* since the latter is pointing to a different city with a different university.

4.3 Allocation to collection categories

In several cases, single units are not recorded due to only one or very few relevant addresses appearing in the database (e.g. hospitals, business enterprises

[11] cf. Winterhager et al., ibid., p. 16ff

or private addresses). For these units, collection categories have been created. An allocation to these collection categories enables to determine if an address belongs to e.g. 'any business enterprise' or 'any hospital' although the single unit is not recorded and therefore no allocation to a single research institution is available.

4.4 Allocation to residual categories

Residual categories are units that were developed for addresses which can be allocated on a higher level of aggregation (e.g. sector or country) but not on the level of the main institution.

The allocation to residual categories occurs via regular expressions as well. An allocation to a sector residual category, however, only occurs in those cases where no concrete allocation to a research institution within the respective sector is possible. The allocation to a sector-residual-category and a concrete research institution of another sector at the same time is, however, possible in individual cases (in the case of more than one research institution mentioned in an address).

Another residual category identifies addresses which are 'German addresses' according to country attribute but which cannot be allocated further (neither to a concrete research institution nor a sector-residual category). Examples are erroneous/incomplete addresses (e.g. addresses consisting of one city or a postal address where the research institution remains unclear). The address

Forschungsinst, D-60325 Frankfurt, Germany

can serve as an example for these kind of addresses.

An address with the allocation to the sector-residual category Max-Planck Society serves as an example:

MPI, MUNCHEN, DEU.

Since there is more than one Max-Planck Institute in Munich, an allocation to a concrete research institution is not possible, whereas the allocation to the sector is.

4.5 Allocation on the lowest hierarchical level

In many cases there are allocations for one address to different hierarchical levels of a research institution. In order to be able to allocate as precisely as possible (i.e. to determine and take into account structural changes and hierarchical relationships as specifically as possible for the concrete address), in each relevant branch of the hierarchy only that unit is taken into consideration that belongs to the lowest possible hierarchical level. There can, however, still remain one or several allocated units per address.

In the following, we will present two examples which, on the one hand, show that an allocation on hierarchical levels lower than that of the main institution can be necessary, and, on the other, why it is useful to allocate exclusively on the lowest hierarchical level and to allocate the respective sub-unit to its main institution(s) in the course of the aggregation (which will be described below).

4.5.1 Example: A unit with two main institutions.

The Ernst Ruska-Centre for Microscopy and Spectroscopy with Electrons (ER-C) is a sub-unit of two main institutions: the Forschungszentrum Jülich and the RWTH Aachen[12] (figure 4).

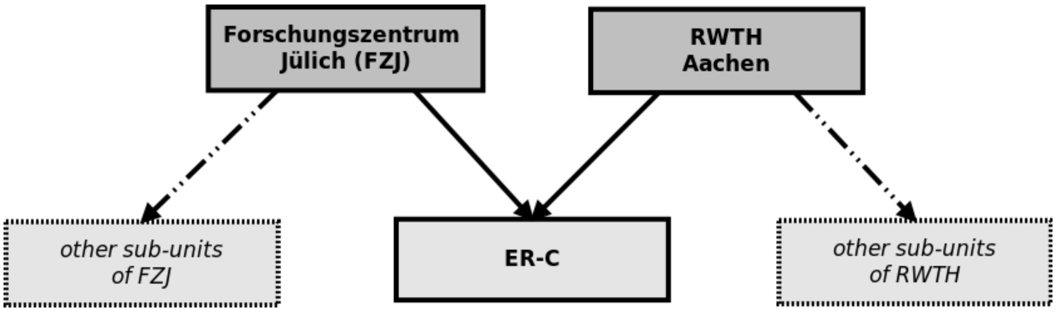


Figure 4: Hierarchical relationships of Ernst-Ruska-Zentrum (ER-C).

In the pattern allocation, the address

Forschungszentrum Julich, ER C Ernst Ruska Ctr, D- 52425 Julich, Germany

receives allocations to the ER-C and to Forschungszentrum Jülich. If there were no allocation to the sub-unit ER-C, or if this were deleted, since it is a sub-unit, only the allocation to Forschungszentrum Jülich would remain. However, an allocation to both the Forschungszentrum Jülich and RWTH Aachen is correct here since the ER-C is a sub-unit of both institutions. If an allocation occurs on the lowest possible hierarchical level (here ER-C), then one obtains the desired result in the aggregation process via the information that the ER-C is a sub-unit of both institutions.

4.5.2 Example: Outsourcing.

Why an exclusive allocation on the lowest possible hierarchical level makes sense can be shown by the examples of the Forschungsgesellschaft für Angewandte Naturwissenschaften e. V. (FGAN) and the Fraunhofer Institute for

[12] cf. URL: http://www.er-c.org/centre/centre.htm (accessed 14.07.2017).

Communication, Information Processing and Ergonomics (FKIE), displayed in figure 5. Until 16 August 2009, the FKIE was sub-unit of the FGAN and then became a main institution itself (outsourcing).

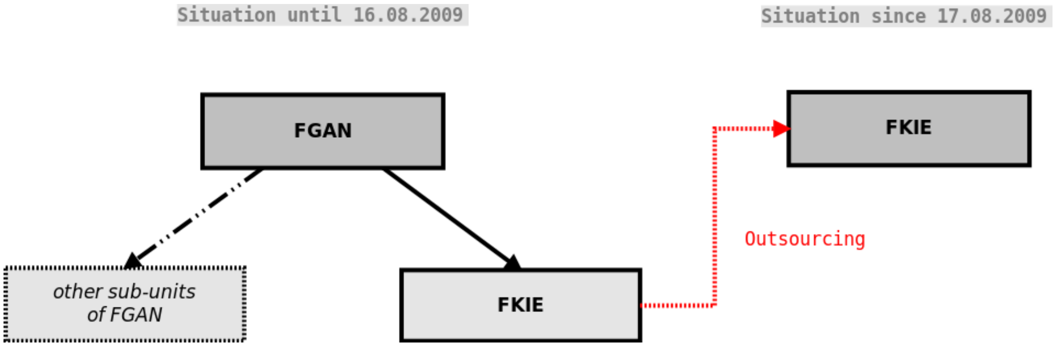


Figure 5: Hierarchical relationships of FKIE.

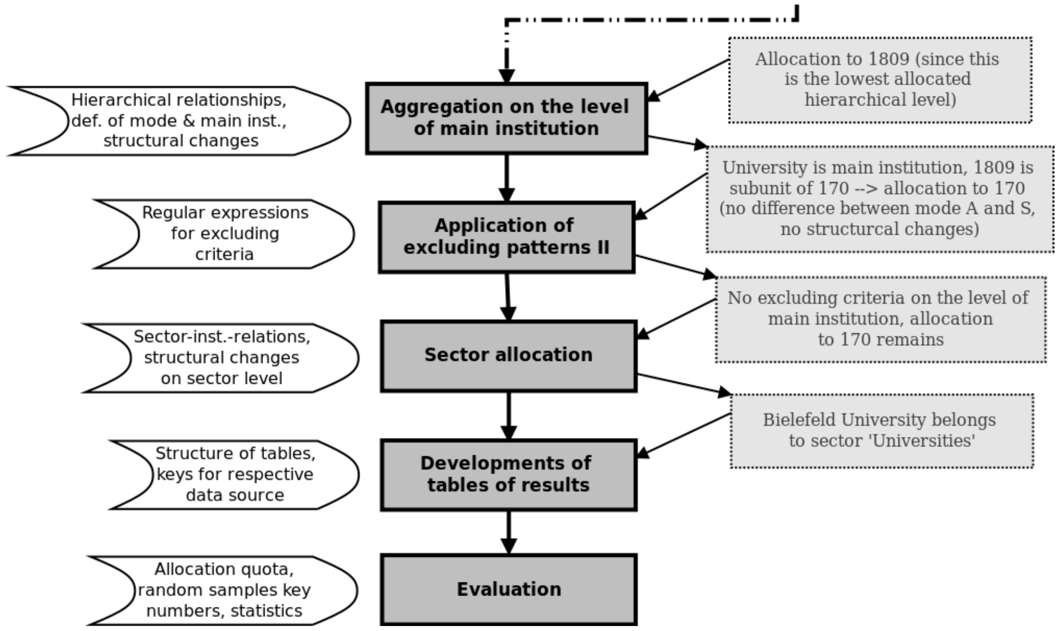
The address

FGAN FKIE, Neuenahrer Str 20, D-53343 Wachtberg, Germany

can be allocated to the FGAN as well as the FKIE via the pattern application. Since the FKIE has a sub-unit relationship with the FGAN, the lowest possible hierarchical level here is the FKIE.
If the allocation here is not carried out on the lowest possible hierarchical level, then both allocations (to FGAN and FKIE) remain. Until 2009 this is unproblematic: the publications on the FKIE are allocated to the main institutional level of the FGAN in the course of the aggregation, since the FKIE is a sub-unit of the FGAN.
This is different with regard to publications after 2009: if both allocations remain, the result is an – unwanted – allocation to the main institution FGAN after its shut-down (follow-up phase) and an allocation to the FKIE, since this is a main institution itself in the year of publication. The continuation of both allocations thus implies an 'artificial cooperation' between an institution that has already been shut-down and its outsourced sub-unit, which is not useful (in particular because neither of the two units were main institutions at the same time, and thus a cooperation between main institutions is not possible).

5 Aggregation and result processing

Figure 6 shows the progress of the procedural steps that follow the application of pattern recognition. Analogous to figures 2 and 3, the necessary inputs, resp. prerequisites, for the steps (shown in the middle) are shown on the left side, while the right side depicts the effects on the example.



Hierarchical relationships, def. of mode & main inst., structural changes
Aggregation on the level of main institution
Allocation to 1809 (since this is the lowest allocated hierarchical level)
University is main institution, 1809 is subunit of 170 --> allocation to 170 (no difference between mode A and S, no structurcal changes)
Regular expressions for excluding criteria
Application of excluding patterns II
No excluding criteria on the level of main institution, allocation to 170 remains
Sector-inst.-relations, structural changes on sector level
Sector allocation
Bielefeld University belongs to sector 'Universities'
Structure of tables, keys for respective data source
Developments of tables of results
Allocation quota, random samples key numbers, statistics
Evaluation

Figure 6: Aggregation and result processing.

5.1 Aggregation on the level of the main institution

In the aggregation, the allocations (that were carried out on the lowest hierarchical level) are assembled up to the level of the main institutions. For this purpose, definitions are necessary regarding the ‘main institution’ as well as on how structural changes are handled over time. Project-specific aspects can play a role for both. For example, it is conceivable that in one project university hospitals should be considered as sub-units of universities (not as main institutions), while in another context publications of the university hospitals should be listed separately from those of the universities (the hospitals thus receive the status of a main institution). Basically, an organization is viewed as a main institution if it is at the top of the identified hierarchical relationships. The definition of the highest hierarchical level is especially oriented towards the criteria of (legal) independence and can be easily determined in most sectors of the German research system[13].

Depending on the context of application, different requirements can also emerge for handling of structural changes over time. Thus, in one case the reference to the historical structure of the institutional landscape in the respective year of publication can be of priority, while in another case all allocations should refer to the current structure (at the time of the analysis). Whether certain institutions, in the case of shutting down, should 'pass' their publications on to a corresponding successor or not depends on such definitions. A prominent example where this is of high relevance is the Karlsruhe Institute for Technology (KIT) which emerged from the fusion of Karlsruhe University and the Forschungszentrum Karlsruhe (FZK).

In order to take these conditions into account, two variations of allocation with different modes are developed: Mode A (allocation according to the current institutional situation) and mode S (consideration of the historical situation at the time of the respective publication)[14].

5.2 Sector allocation

In addition to the aggregation on the level of the main institution, an allocation is carried out on the sector level. For many projects, this allocation to the large sectors of Germany's system of science (universities, universities of applied sciences, FhG, HGF, MPG, WGL, etc.) represents an important resource. Changes of the sector allocation over time are identified separately and taken into account in the mentioned modes (A and S) accordingly.

5.3 Making the results available

The results are listed in a set of tables for both modes of allocation. These tables are related to one another via corresponding keys[15]. The tables with the allocation of addresses to institutions (and those with the allocation of addresses to sectors) form the core. These tables are developed and presented in several variations: one each per mode of allocation (A vs. S) and per source of the address data sets (currently: Web of Science and Scopus). They contain the links (by means of corresponding identifiers) of the publication and address data sets with the fitting data sets of the tables for the institutions and sectors. The latter two serve as look-up tables and contain detailed information on the identified main institutions (name, start and end date) and sectors. An additional table is provided for the mode of the synchronic allocation which lists

[13] For details on the definition of the main institution, cf. Winterhager et al., ibid., p. 7.

[14] For details on these modes, cf. Winterhager et al., ibid., p. 9

[15] A technical documentation which provides details on the structure of these tables (which are also made available for the Competence Centre for Bibliometrics) is available on request.

the predecessor-successor relationships on the level of the main institutions with the date of the transition[16]. It can be used as a basis in mode S in order to realize project-specific variations with individual deviations, resp. adaptations. The allocation of institutional sectors is independent of the database and listed in one table per mode each.

5.4 Evaluation

For the quality control of the procedure as well as for obtaining information about previously unidentified units and necessary changes/updates in the pattern inventory, allocation quotes are calculated after each coding phase, plausibility tests and random samples (drawn from the allocated and non-allocated addresses) are conducted and checked manually. The insights gained in the evaluation can help to add, correct and modify the basic data, which can then be used in an updated form in the next coding.

Examples for additions/corrections are the inclusion of previously unidentified units, deletion of patterns that create false hits, addition of new patterns for newly identified units or previously non-allocated addresses of already identified units, identification of further sub-units of a unit, identification of further structural changes and relationship between units.

Aside from an evaluation of the contents, the procedure is constantly tested as well. Finally, figure 7 shows all steps of the procedure and their connection.

[16] This table is not necessary for mode A because the structural changes have already been taken into account here.

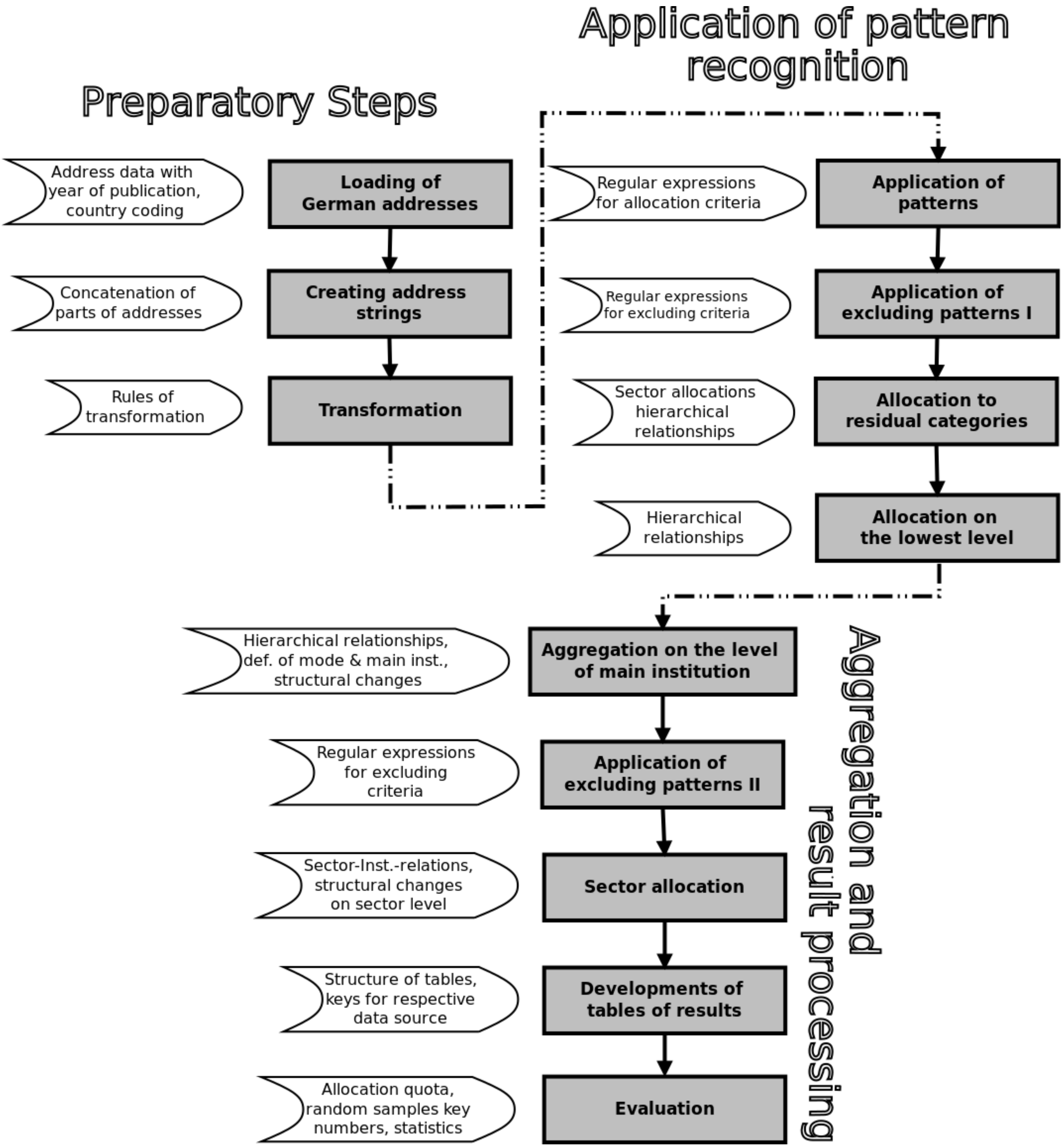

Preparatory Steps
Application of pattern recognition
Aggregation and result processing
Address data with year of publication, country coding
Loading of German addresses
Concatenation of parts of addresses
Creating address strings
Rules of transformation
Transformation
Regular expressions for allocation criteria
Application of patterns
Regular expressions for excluding criteria
Application of excluding patterns I
Sector allocations hierarchical relationships
Allocation to residual categories
Hierarchical relationships
Allocation on the lowest level
Hierarchical relationships, def. of mode & main inst., structural changes
Aggregation on the level of main institution
Regular expressions for excluding criteria
Application of excluding patterns II
Sector-Inst.-relations, structural changes on sector level
Sector allocation
Structure of tables, keys for respective data source
Developments of tables of results
Allocation quota, random samples key numbers, statistics
Evaluation

Figure 7: Processing steps of the disambiguation (summary).

6 Statistics

6.1 Basic data

In July 2017 about 4,600 units with the corresponding attributes are recorded on different hierarchical levels (including 9 residual categories), 2,300 of which have been main institutions at some point in time of their existence. For these units, approximately 51,600 patterns are recorded and form the basis of the disambiguation procedure.

6.2 Precision & recall

Precision and recall of the disambiguation process depends on several factors including quality and pre-standardization of source data (e.g. number of incomplete addresses, number of addresses mentioning only sub-units, number of variations per institution).

Recall can be defined in different ways in this context, e.g. the share of document-address-combinations assigned, the share of distinct author addresses assigned or the share of publications with assignment.
Table 3 shows recall values corresponding to different definitions for Web of Science data from May 2017 (German addresses), publication years 1980 to 2017.

number of distinct addresses	2,042,927
number of distinct addresses with at least one assignment	1,761,595
share of dist. addresses with at least one assignment	**0.8623**
number of (distinct) document-address-combinations	6,056,563
number of (dist.) document-address-combinations, assigned	5,646,488
share of (dist.) doc.-addr.-combinations, assigned	**0.9323**
number of documents	3,352,370
number of documents with at least one assignment	3,169,871
share of dcuments with at least one assignment	**0.9456**

Table 3: Recall for Web of Science data, 1980-2017.

Thereof, 8,641 document-address-combinations have been allocated to residual categories, including 1,106 classified as impossible to assign (due to incomplete or erroneous address data) – and therefore allocated to the residual category 'German addresses' containing addresses without the possibility of more precise allocations (e.g. to sectors or research institutions).

As apparent in table 3, recall values are lower for distinct addresses. This is an effect of the longtail of addresses appearing only once or with very low frequency in the data sources.

Precision has been evaluated via a manual check of a random sample (1,000 assignments of document-address-combinations to institutions, again Web of Science data from May 2017, year of publication 1980-2017, following mode S described above), resulting in a share of 99.3% allocations processed correctly, where in 0.6% (6 allocations) errors arise due to incomplete recording of structural changes (document-address-combinations with a year of publication before a fusion have been assigned to the successor – which is correct in mode A but not in mode S).