

Discourse Network Analyzer

Philip Leifeld and Johannes Gruber

Last update: DNA 2.0 beta 19 with rDNA 2.0.1 on 2018-01-11.

This document describes the software **Discourse Network Analyzer** (DNA) and its companion R package **rDNA** (Leifeld 2017). Section 1 is a concise and fairly technical description of the types of networks DNA can export. Section 2 explains how to install both DNA and rDNA which both rely on a correctly set up Java runtime environment. Then two sections follow, which describe the usage of DNA in detail: Section 3 describes how to set up a project in DNA, including the creation of a database, adding and managing users and describes how to set up statement types and variables. Section 4 explains how you import and organise your raw data (i.e. documents). Section 7 is an introductory tutorial on using the rDNA package. The contents of both sections are meant to be extended in the future, and additional sections may be added. DNA and rDNA can be downloaded from [GitHub](#). Questions and bug reports can be posted in the issue tracker on [GitHub](#).

1 DNA algorithms

This section summarizes the main algorithms implemented in DNA in a technical way.

X is a three-dimensional array representing statement counts. x_{ijk} is a specific count value in this array, with the first index i denoting an instance of the first variable (e.g., organization i), the second index j denoting an instance of the second variable (e.g., concept j), and the third index k denoting a level on the qualifier variable (e.g., agreement = 1). For example, $x_{ijk} = 5$ could mean that organization i mentions concept j with intensity k five times.

Where the qualifier variable is binary, *false* values are represented as 0 and *true* values as 1 on the k index, i.e., $K^{\text{binary}} = \{0; 1\}$. Where the qualifier variable is integer, the respective integer value is used as the level. This implies that k can take positive or negative values or 0, i.e., $K^{\text{integer}} \subseteq \mathbb{Z}$. Note that all k levels of the scale are included in K , not just those values that are empirically observed.

Indices with a prime denote a second instance of an element, e.g., i' may denote another organization. Y denotes the output matrix to be obtained by applying a transformation to X . The following transformations are possible:

1.1 Congruence

In a congruence network, the edge weight between nodes i and i' represents the number of times they co-support or co-reject second-variable nodes (if a binary qualifier is used) or the cumulative similarity between i and i' over their assessments of second-variable nodes (in the case of an integer qualifier variable).

In the integer case:

$$y_{ii'}^{\text{congruence}} = \Phi_{ii'} \left(\sum_{j=1}^n \sum_k \sum_{k'} x_{ijk} x_{i'jk'} \left(1 - \frac{|k - k'|}{|K| - 1} \right) \right) \quad (1)$$

where $\Phi_{ii'}(\cdot)$ denotes a normalization function (to be specified below).

In the binary case, i.e., $|K| = 2$, this reduces to

$$y_{ii'}^{\text{congruence binary}} = \Phi_{ii'} \left(\sum_{j=1}^n \sum_k x_{ijk} x_{i'jk} + (1 - x_{ijk})(1 - x_{i'jk}) \right). \quad (2)$$

1.2 Conflict

Binary case:

$$y_{ii'}^{\text{conflict binary}} = \Phi_{ii'} \left(\sum_{j=1}^n \sum_k (1 - x_{ijk}) x_{i'jk} + x_{ijk} (1 - x_{i'jk}) \right). \quad (3)$$

More generally, in the integer case:

$$y_{ii'}^{\text{conflict}} = \Phi_{ii'} \left(\sum_{j=1}^n \sum_k \sum_{k'} x_{ijk} x_{i'jk'} \left(\frac{|k - k'|}{|K| - 1} \right) \right) \quad (4)$$

1.3 Subtract

$$y_{ii'}^{\text{subtract}} = y_{ii'}^{\text{congruence}} - y_{ii'}^{\text{conflict}} \quad (5)$$

1.4 Ignore

$$y_{ii'}^{\text{ignore}} = \Phi_{ii'} \left(\sum_{j=1}^n \left(\left(\sum_k x_{ijk} \right) \left(\sum_k x_{i'jk} \right) \right) \right) \quad (6)$$

1.5 Normalization

In the simplest case, normalization can be switched off, in which case $\Phi_{ii'}^{\text{no}}(\omega) = \omega$.

Alternatively, edge weights can be divided by the average activity of nodes i and i' :

$$\Phi_{ii'}^{\text{avg}}(\omega) = \frac{\omega}{\frac{1}{2} \left(\sum_{j=1}^n \sum_k x_{ijk} + \sum_{j=1}^n \sum_k x_{i'jk} \right)}. \quad (7)$$

With Jaccard normalization, we don't just count i 's and i' 's activity and sum them up independently, but we add up both their independent activities and their joint activity, i.e., both matches and non-matches:

$$\Phi_{ii'}^{\text{Jaccard}}(\omega) = \frac{\omega}{\sum_{j=1}^n \sum_k x_{ijk} [x_{i'jk} = 0] + \sum_{j=1}^n \sum_k x_{i'jk} [x_{ijk} = 0] + \sum_{j=1}^n \sum_k x_{ijk} x_{i'jk}}. \quad (8)$$

With cosine normalization, we take the product in the denominator:

$$\Phi_{ii'}^{\text{cosine}}(\omega) = \frac{\omega}{\sqrt{(\sum_{j=1}^n \sum_k x_{ijk})^2} \sqrt{(\sum_{j=1}^n \sum_k x_{i'jk})^2}}. \quad (9)$$

1.6 Affiliation networks

Ignoring the qualifier variable:

$$y_{ij}^{\text{affiliation ignore}} = \Phi_{ij} \left(\sum_k x_{ijk} \right) \quad (10)$$

Subtracting negative from positive ties (integer case):

$$y_{ij}^{\text{affiliation subtract binary}} = \Phi_{ij} \left(\sum_k k \cdot x_{ijk} \right) \quad (11)$$

Subtracting negative from positive ties (binary case):

$$y_{ij}^{\text{affiliation subtract binary}} = \Phi_{ij} \left(\sum_k (k \cdot x_{ijk} - (1 - k) \cdot x_{ijk}) \right) \quad (12)$$

Note that the binary case is *not* merely a special case of the weighted affiliation network in this case.

1.7 Normalization for affiliation networks

With *activity* normalization, ties from active nodes receive lower weights:

$$\Phi_{ij}^{\text{activity}}(\omega) = \frac{\omega}{\sum_{j=1}^n \sum_k x_{ijk}} \quad (13)$$

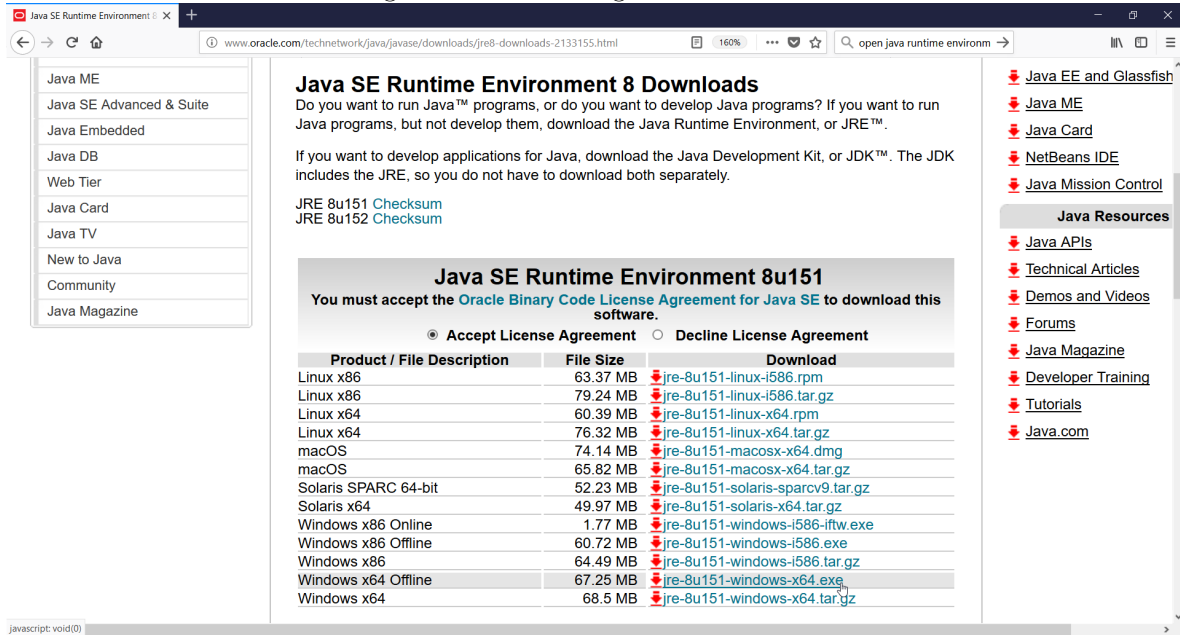
With *prominence* normalization, ties to prominent nodes receive lower weights:

$$\Phi_{ij}^{\text{prominence}}(\omega) = \frac{\omega}{\sum_{i=1}^m \sum_k x_{ijk}} \quad (14)$$

2 Installation of DNA and rDNA

This section explains how DNA and rDNA can be installed on common desktop operating systems. As DNA is written in Java, both DNA and rDNA rely on a Java to work on your computer properly. Installing and configuring a valid Java Runtime Environment on your machine will thus be the first and only complicated step of the installation. However, following the simple steps below, one should not run into problems while setting up Java. The advantage of the Java programming language for academic software is that it both runs on different operating systems without altering the source code—once the Runtime Environment is set up—and that it is—for the most part—open source. Besides setting up the Java Runtime Environment, installation of DNA and rDNA is identical on different operating systems. If you feel confident that Java is already correctly set up on your computer, you can therefore skip to section subsection:installdna. Otherwise please continue to the section for the operating system you wish to install DNA and rDNA on: [Windows](#), [macOS](#) or [Linux](#).

Figure 1: Downloading JRE from Oracle



2.1 Windows

To install the necessary Java Runtime Environment on your Windows computer, simply go to oracle.com/technetwork/java/javase/downloads/, scroll down to **"JRE DOWNLOAD"** (JRE being the acronym for Java Runtime Environment) and then scroll down to the newest version of your respective Windows version. In this example, I'm running "Windows 10 Home 64 Bit" which means that we need to download and install **"jre-8u151-windows-x64.exe"** (see figure 1). During the installation, you can accept all the default options, including the installation path.

Next, you should set **"JAVA_HOME"** in your environmental variables to tell your Windows PC where your Java installation lives. This step is optional but can prevent many issues with Java people had in the past. To set **"JAVA_HOME"**, you need to navigate to the menu **"edit the system environment variables"**. The easiest way to get there is to hit the **"Windows Button"** on your keyboard and enter **"environment"**. Windows will then search for programs and settings menus which include this title and should usually display the menu we are looking for on top.¹ In this menu you have to find the button **"Environment variables..."**. Clicking this button should open the windows shown in figure 2.

Under User Variables, click New.² Enter the variable name **"JAVA_HOME"** and the path to your java installation in the field **"Variable value"**. If you haven't altered the default install location, you should find Java in **"C:\Program Files\Java\jre1.8.0_151"** or in **"C:\Program Files (x86)\Java\jre1.8.0_151"** if you chose to install a 32bit version of Java (which might cause problems if you try to use it with the 64bit version of R).

Windows should now recognise Java and be able to run Java-commands. To test this, we can open the

¹On older versions of Windows, this might not work. On Windows 7 you can alternatively right-click on "My Computer" and select "Properties → Advanced". On Windows 8 "Control Panel → System → Advanced System Settings".

²This sets **"JAVA_HOME"** just for the current user. If you want to make Java available for all users you can enter a System Variable instead.

Figure 2: Edit JAVA_HOME to tell Windows where your Java lives

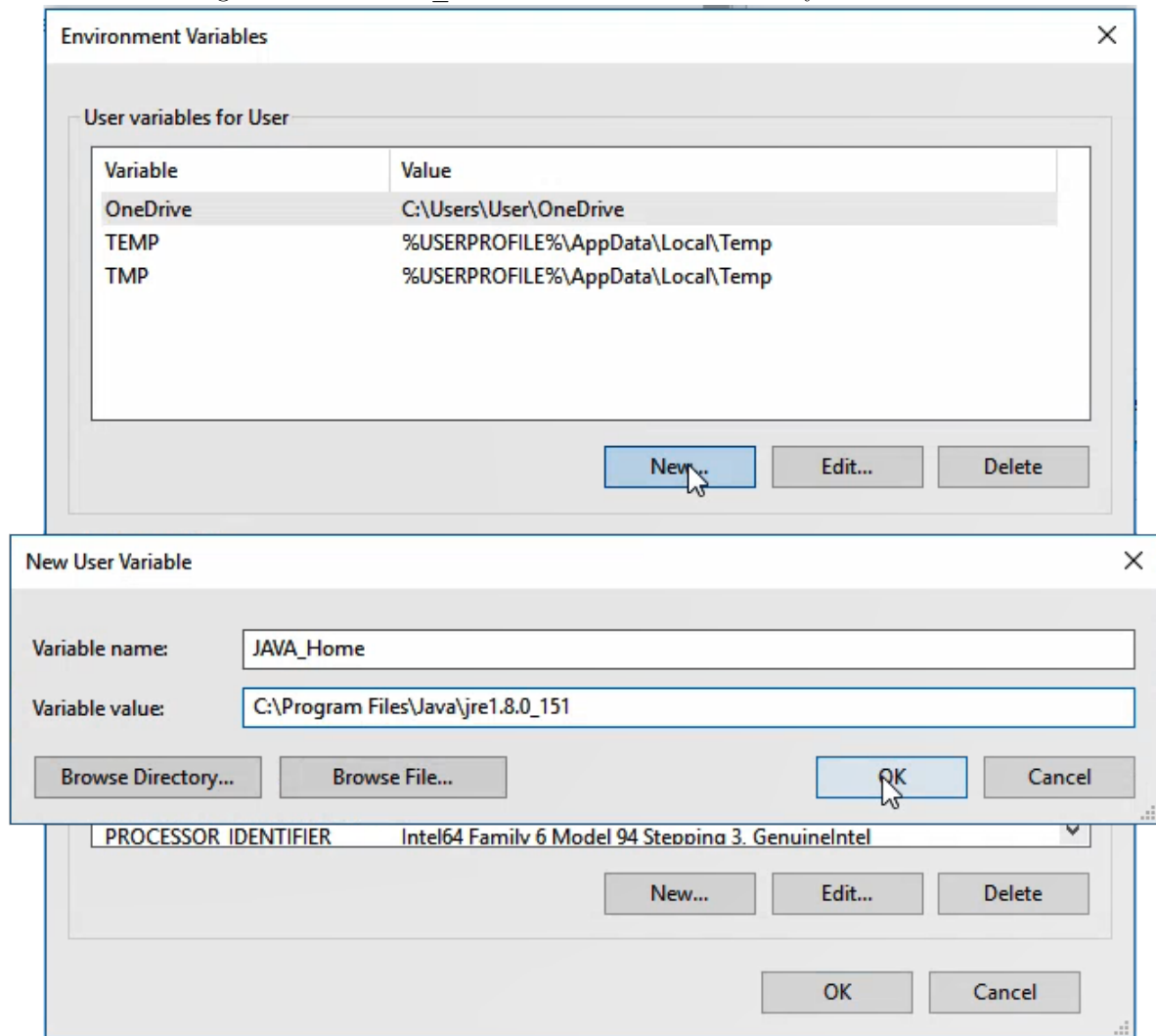
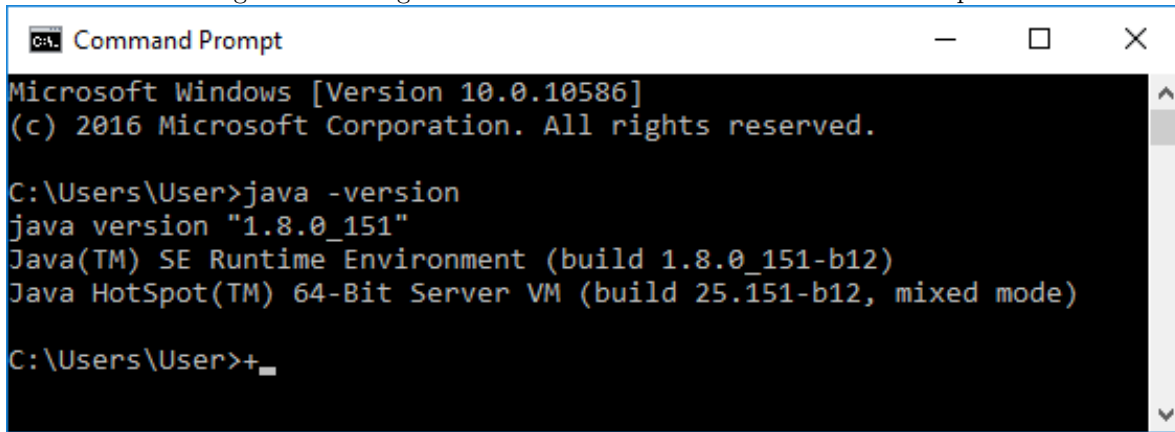


Figure 3: Testing Java installation in Windows Command Prompt

A screenshot of the Windows Command Prompt window. The title bar says "Command Prompt". The text inside shows the Windows version and copyright information, followed by the command "java -version" and its output: "java version '1.8.0_151'", "Java(TM) SE Runtime Environment (build 1.8.0_151-b12)", and "Java HotSpot(TM) 64-Bit Server VM (build 25.151-b12, mixed mode)". The prompt "C:\Users\User>+" is visible at the bottom.

```
Microsoft Windows [Version 10.0.10586]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\User>java -version
java version "1.8.0_151"
Java(TM) SE Runtime Environment (build 1.8.0_151-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.151-b12, mixed mode)

C:\Users\User>+_
```

Command Prompt (press the "Windows Button" on your keyboard and simply enter "cmd" and then hit "Enter") and typing a Java command, e.g. "java -version". If the installation was successful, the output should display information about the Java-version and build as depicted in figure 3.

After installing Java, you are ready to use rDNA and could skip to section 2.4 if you are not interested in installing rDNA as well. In order to use rDNA the rest of this section will explain how to install R and a recommended **integrated development environment (IDE)** called **RStudio**, which makes working with R a lot easier and also looks a lot better than the default GUI.

Install R on Windows

1. First, you need to download R from cran.r-project.org/bin/windows/base/.
2. On the top of the page click on **Download R 3.4.3 for Windows** (or a newer version if available).
3. Install the download file, e.g. "R-3.4.3-win.exe". Usually, it is fine to leave all default settings in the installation options.
4. Go to rstudio.com/products/rstudio/download/.
5. At the bottom of the page, under "Installers for Supported Platforms", click on the link **RStudio 1.1.383 - Windows Vista/7/8/10** (or a newer version if available). Again the default installation options are fine in most cases and can be accepted unchanged.
6. After installation, you can use R by opening RStudio.


Traditionally, the first test you perform in a new programming language is to write a "Hello, World!" program. To do this in R you simply type `print("Hello World!")` in the "Console". Alternatively, you can make R perform a simple mathematical operation. If everything is set up correctly, the output should look like this:

```
print("Hello World!")

## [1] "Hello World!"
```

```
2*3
```

```
## [1] 6
```

The chunk of code above marks the first time we are using R commands in this manual, so it might make sense to explain what this means for users who are not familiar with dynamic reports which contain R-code. Whenever R-code is shown in this manual it is decorated with a light grey background. The output which is generated by running a command is marked by two #, which means that every line which does not start with ## contains R-code which you can copy & paste to the Console in RStudio and run. You can either run a command in RStudio by clicking on this button  Run on the upper right of RStudio, near the corner.

2.2 macOS

WIP

After installing Java, you are ready to use DNA and could skip to section 2.4 if you are not interested in installing rDNA as well. In order to use rDNA the rest of this section will explain how to install R and a recommended integrated development environment (IDE) called RStudio, which makes working with R a lot easier and also looks a lot better than the default GUI. Look for the steps below to

install R on Mac

- 1.

Traditionally, the first test you perform in a new programming language is to write a "Hello, World!" program. To do this in R you simply type `print("Hello World!")` in the "Console". Alternatively, you can make R perform a simple mathematical operation. If everything is set up correctly, the output should look like this:

```
print("Hello World!")
```

```
## [1] "Hello World!"
```

```
2*3
```

```
## [1] 6
```

2.3 Linux

WIP

After installing Java, you are ready to use DNA and could skip to section 2.4 if you are not interested in installing rDNA as well. In order to use rDNA the rest of this section will explain how to install R and a recommended integrated development environment (IDE) called RStudio, which makes working with R a lot easier and also looks a lot better than the default GUI.

install R on Linux

1.

Traditionally, the first test you perform in a new programming language is to write a "Hello, World!" program. To do this in R you simply type `print("Hello World!")` in the "Console". Alternatively, you can make R perform a simple mathematical operation. If everything is set up correctly, the output should look like this:

```
print("Hello World!")

## [1] "Hello World!"

2*3

## [1] 6
```

2.4 Installing the programs themselves

Once Java is set up correctly, you can simply download the latest version of DNA as a JAR file from github.com/leifeld/dna/releases (see figure 4). JAR or .jar files are technically archive files which usually contain a computer program written in Java, along with all the pictures and libraries necessary to run the program. Once the download is finished, you can start the program by double-clicking on the downloaded file. However, on Linux, it is sometimes necessary to make the file executable first (e.g. via `$chmod +x /path/to/your/dna.jar` or using a GUI-method). If clicking on the file does not open the program on a windows machine, right-click on the .jar file → "Open with" → "Use another app" and then navigate to the file "C:\Program Files\Java\jre1.8.0_151\bin\javaw.exe". If you are not interested in using rDNA, you can now skip to the next section.

At this point, I assume that you have installed R and have at least a minimal understandint of how the program works. If that is not the case, you might want to jump back to where we explain how to install [Install R on Windows](#), [install R on Mac](#) or [install R on Linux](#). If you have already done this, we can go ahead and install rDNA from within R. First, we need to install the package rJava (Urbanek 2016), which is the most important dependency of rDNA:

```
install.packages("rJava")
```

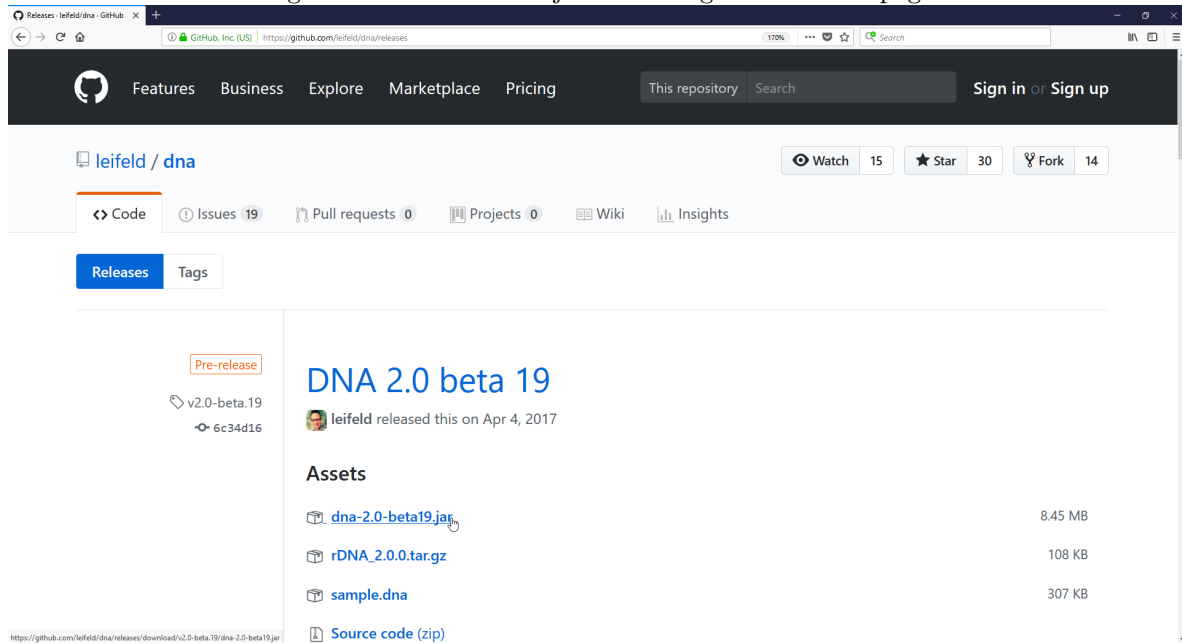
To see if this worked, or to troubleshoot potential problems, we can run a couple of Java-commands from within R:

```
library(rJava)
# 1. initialize JVM
.jinit()
# 2. retrieve the Java-version
.jcall("java/lang/System", "S", "getProperty", "java.version")

## [1] "1.8.0_151"

## [1] "1.8.0_151"
```


Figure 4: Download DNA jar file from github releases page



```
# 3. retrieve JAVA_HOME location
.jcall("java/lang/System", "S", "getProperty", "java.home")

## [1] "/usr/lib/jvm/java-8-openjdk-amd64/jre"

## [1] "C:\\Program Files\\Java\\jre1.8.0_151"

# 4. retrieve Java architecture
.jcall("java/lang/System", "S", "getProperty", "sun.arch.data.model")

## [1] "64"

## [1] "64"

# 5. retrieve architecture of OS
.jcall("java/lang/System", "S", "getProperty", "os.arch")

## [1] "amd64"

## [1] "amd64"

# 6. retrieve architecture of R as well
R.Version()$arch

## [1] "x86_64"

## [1] "x86_64"
```

Now what you want to make sure, in case something is not working correctly with `rJava`, is if the architectures of `Java` *see step 4*, your operating system *see step 5* and your version of `°` *see step 6* match.

Once this is done, you should install the package `devtools` (Wickham and Chang 2016), which permits installing R packages from `GitHub`.

```
install.packages("devtools")
```

Since we only need one function from the package `devtools` at this point, it is not necessary to invoke the `library` command to load the whole package. Instead you can write `"devtools::"` and then type the function you want to use.³

```
devtools::install_github("leifeld/dna/rDNA",  
  args = "--no-multiarch")
```

After this is done as well, the final step of the installation is to test if `rDNA` can be loaded correctly and to perform a basic operation with it—opening DNA from within R. In order to do so, you first need to download DNA, which can also be done in R with the `download.file`-command (*see section 7 for more information about this code chunk*).

```
# download two files necessary to test rDNA  
download.file(  
  "https://github.com/leifeld/dna/releases/download/v2.0-beta.19/dna-2.0-beta19.jar",  
  destfile = "dna-2.0-beta19.jar", mode = "wb") # download DNA jar  
download.file(  
  "https://github.com/leifeld/dna/releases/download/v2.0-beta.19/sample.dna",  
  destfile = "sample.dna", mode = "wb") # download sample file  
  
# load library  
library(rDNA)  
  
# initialise the file you just downloaded  
dna_init("dna-2.0-beta19.jar")  
  
# start up DNA from R with the sample file to see if everything worked  
dna_gui(infile = "sample.dna")
```

3 Using DNA:Preparation of your DNA Workspace

After installing the program (see Section ??), you can now create your first DNA database for your own research project. How you set up a DNA database will mainly depend on the needs of your personal research design—which should usually be clear before you start analysing data. Therefore, DNA can be customised during the creation of a new database in accordance with how you are planning to use the tool. Note, however, that several of the decisions made during the initial setup of a database can thus not be altered, except by creating a new database and importing the old codes and data.

³The option `"args = "--no-multiarch"` should normally not be necessary but prevents errors on some operating systems, since `devtools` tries to test both 32-bit and 64-bit version of a package—which inevitably causes an error as only one architecture of `Java` is available.

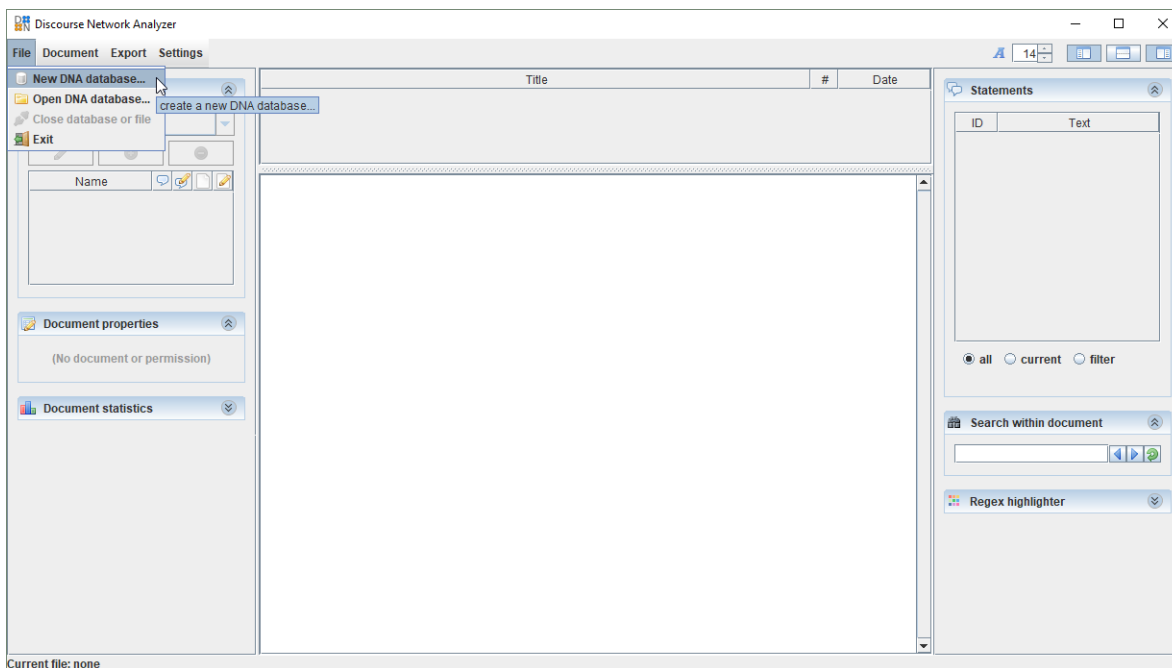


Figure 5: Starting a new Database

3.1 Creating a new DNA database

In order to create a new DNA database-file, you have to click on the index tab **"File"** (in the upper left corner of your DNA program window) and select the option **"New DNA database"** (*cf. figure 5*). As a result, a new window will open (*cf. figure 6*), in which you find a menu that provides you with a step-by-step guidance for specifying the configuration of your personal DNA project.

Clicking on the first tab in the sidebar of this menu—**"Database"** (*cf. figure 6*)—opens a menu, which allows you to choose the file-name and storage location of your database. For this first step of your set-up, DNA provides you with two options in respect to the type of database, in which your data is stored. Which of these options best fits your research project is dependent on the circumstances of your coding process:

The preset option **"Local .dna file"** means, that the dataset is stored in a local file⁴ on your PC or device. This file, with the file extension .dna, can be moved on your machine, sent via email, uploaded and shared via a cloud file hosting service—such as Dropbox—and can generally be treated in the same way as any other file PC users are familiar with. This will be sufficient in most user scenarios, for example, if you employ a single coder working on a single computer, if multiple coders work on a single dataset at non-overlapping intervals or when multiple coders work at the same time on different datasets, which you merge after the coding process (*see section 4.2.3*). For most users, this simpler option will be sufficient and in order to use DNA it is not necessary to be familiar with setting up and managing an SQLite or MySQL-database.

However, for more experienced user or research projects in which several coders want to work on the same database at the same time, a second option was included into DNA **"Remote database on a server"**. This stores your data in a MySQL database which could be stored locally on your machine—which would defy the purpose though—on a

⁴technically an SQLite file.

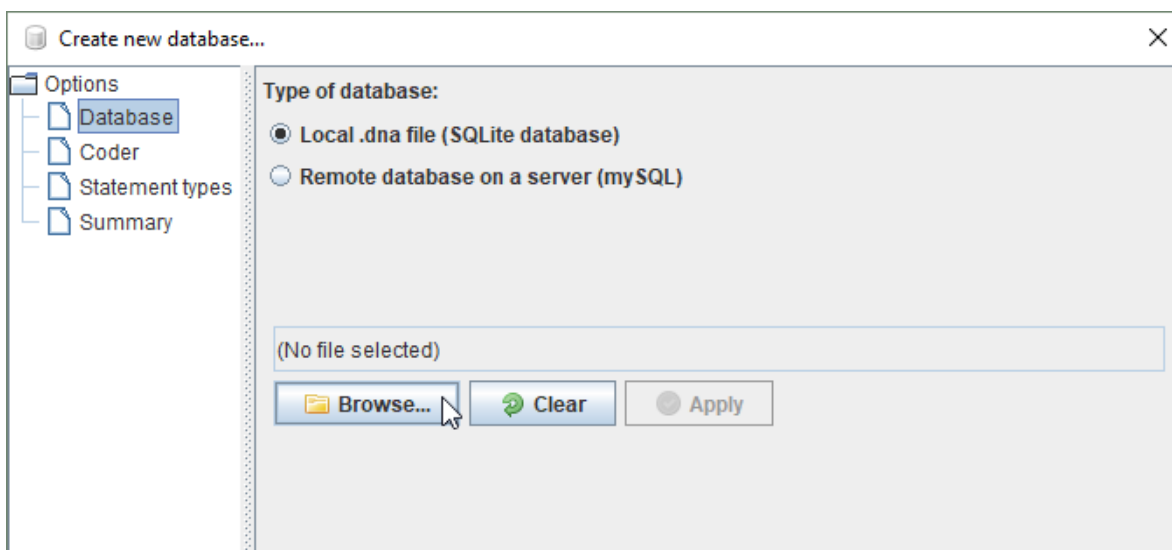


Figure 6: Choose if database will be stored locally or remotly

private sever—such as a Network-attached storage (NAS)—or on an online Cloud server. You should select this option if you employ a single coder working on multiple devices or multiple coders working on a single dataset at the same time. The preconditions for using this type of storage are that all coders have a stable connection to the database during the coding process—e.g. via the internet—and that you [set up an online MySQL database](#) in advance.

3.1.1 Creating a local DNA file

1. Click on the button **"Browse"** (*cf. figure 6*). Now a pop-up menu—similiar to the one shown in *figure 7*—should be opened.
2. In this pop-up menu, you can choose the storage location of your database on your local device from the **"Save in"**-slide down menu, enter the name of your database in the field **"File Name"** and confirm your choices by pressing the **"Save"**-Button (*cf. figure 7*). Now the pop-up menu will close.
3. **Now it is important, that you confirm your choices again by pressing the "Apply"-Button** (*cf. figure 8*). If you forget to press this Button, you cannot create the database in the final step, because the program will report "No database selected" (*cf. figure 18*).

If you just employ a single coder and don't want to change or supplement the preset standard research variables ("person", "organization", "concept", "agreement") or types of codeable statements ("Statement", "Annotation"), you can now proceed directly to the [final step](#) . If you use this manual as a Beginner's Tutorial for working with DNA, however, it would be helpful to follow the steps outlined in sections [3.2](#) and [3.3](#) in order to gain a better understanding of the DNA's potential uses and its functions.

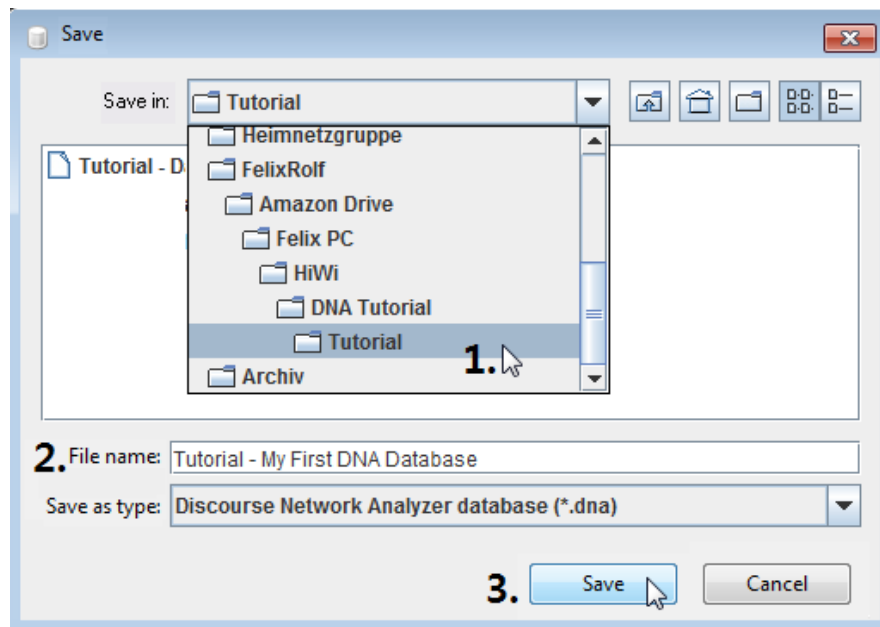


Figure 7: Choose location of database window

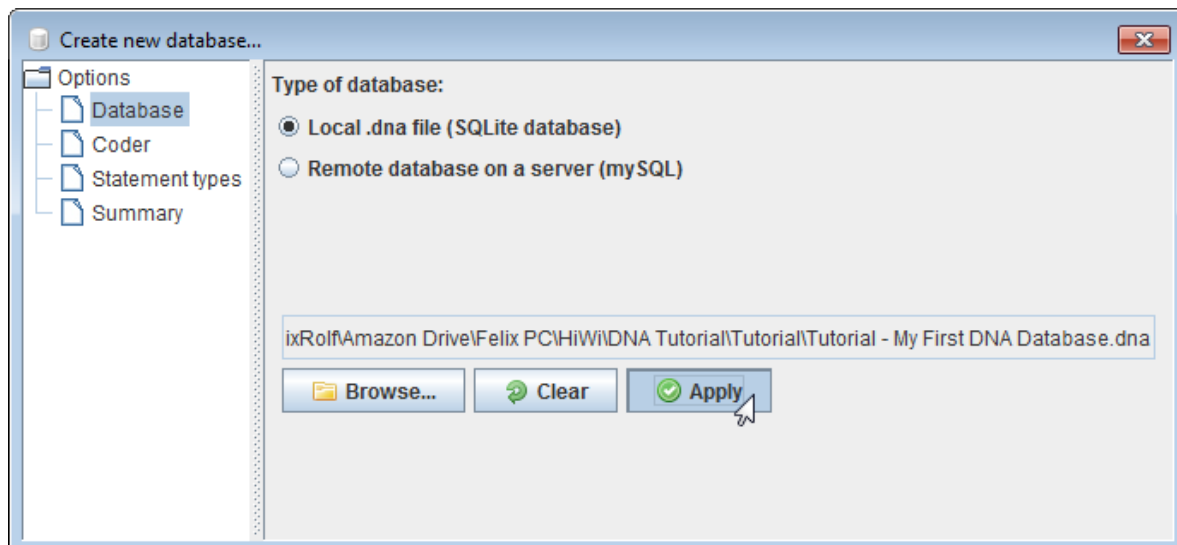


Figure 8: Apply database choice

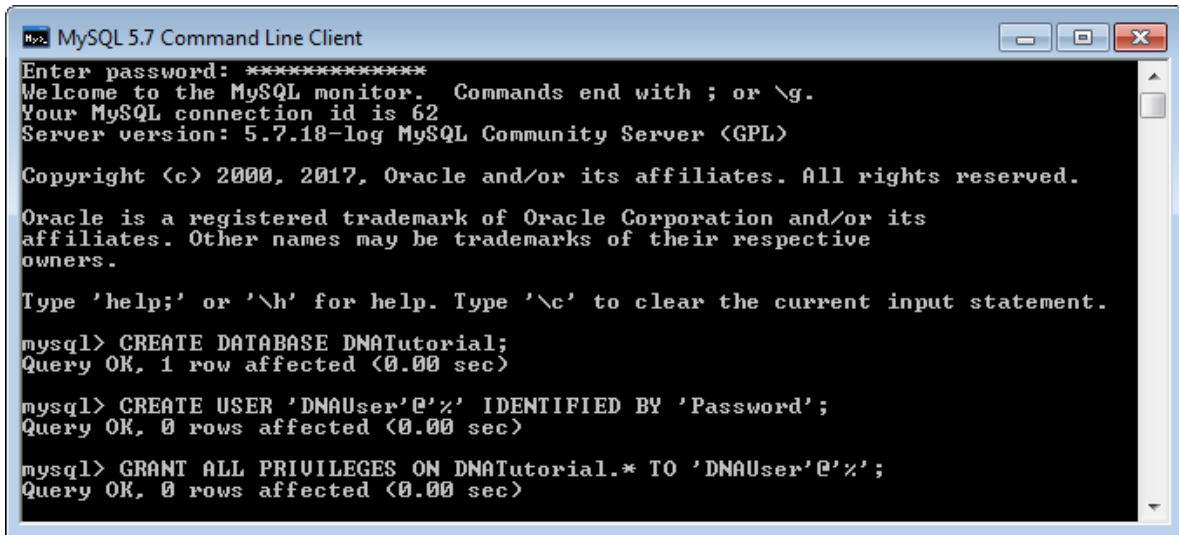


Figure 9: Create MySQL database

3.1.2 Creating and using a remote database (MySQL)

Before you can configure DNA for working with a remote MySQL database, it is necessary to execute at least three basic operations in MySQL (*cf. figure 9*).⁵

1. You have to create a database on your MySQL-server (*usually by the command "CREATE DATABASE 'DatabaseName' "*)
2. As you probably don't want to allow all coders access to all other databases stored on your MySQL-server, you should create distinct user-profile(s) for the coding process of your DNA-project. Even if DNA itself allows for **managing multiple different coder-roles**, it is recommendable to create separate user-profiles for each of the individual coders—especially if they simultaneously edit the content of your database. It is also advisable to create passwords for the access to your database, not only for safety reasons but also because DNA sometimes has problems with signing in users without a password. Consequently you would use the **"CREATE USER 'Username'@'%' IDENTIFIED BY 'Password'"** - Command. It should be noted, that—if necessary—in this step you can restrict the respective users access to your database to a specific device (by replacing '%' through a particular server-address).
3. Finally you, have to equip the users with the necessary rights to edit your database. For this, you use the **"GRANT ALL PRIVILEGES ON Databasename. TO 'Username'@'%'"** -Command, because you can specify distinct user roles and rights with DNA itself in **the next step**, which is particularly designed for discourse network-analytical coding purposes.

Once the MySQL-database is set up, you only have to select the option **"Remote database on a server"** in the first tab of the sidebar menu **"Database"** in DNA (*cf. Creating a new DNA database*) and enter the respective username and password created in the previous step in the respective fields **"User"** and **"Password"** as well as to specify the server-address of the database, with which you want to connect, in the field **"mysql://"**. If you want to access the database remotely from another device, you have to indicate the URL or IP-address of your host-server, the port (which

⁵for a detailed introduction to database management with MySQL *cf.* dev.mysql.com/doc/mysql-getting-started.

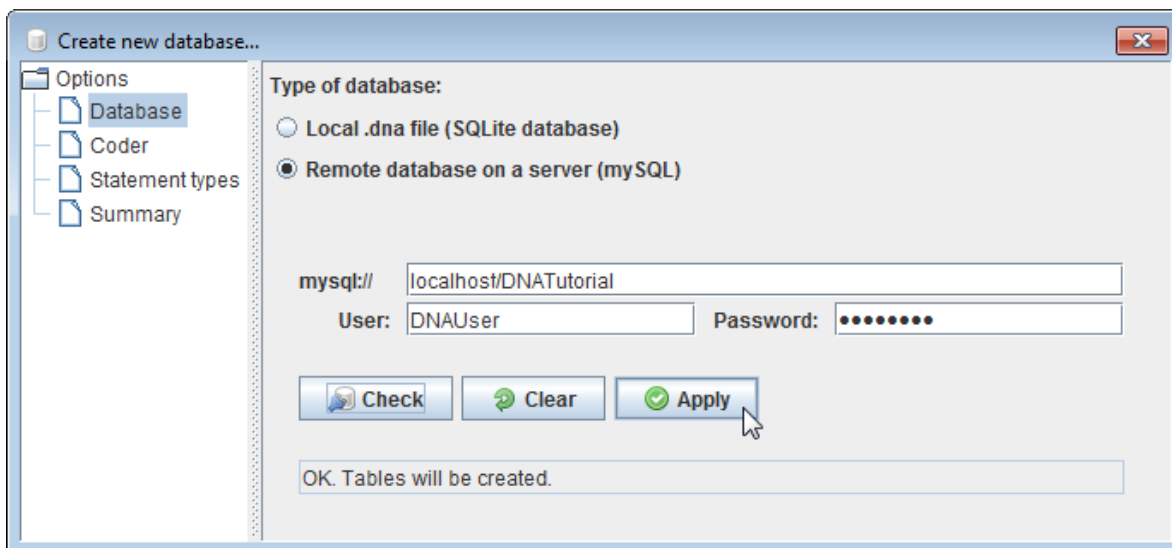


Figure 10: Connecting to local MySQL database

is 3306 in default, but can be **configured manually**) and the name of your database in the format **"Hostserveraddress:Port/Databasename"**. If you use DNA on the device hosting the database you can instead use the configuration shown in *figure 10* (**"localhost"/Databasename**). By clicking the button **"Check"** you can now check if DNA is able to connect to your database. If this is successful, you will receive the message **"Ok. Tables will be created"** (*cf. figure 10*); if not, DNA will report **"Error: Connection could not be established"**. In case of the latter, you should check the validity of your server address, username and password and—if necessary—repeat the steps outlined above. It should be noted that—for security reasons—MySQL doesn't allow remote access with the **"root"**-superuser-profile in most cases. Similar to the generation of a local .dna file, it is finally important, that you confirm your choices again by pressing the **"Apply"-Button** (*cf. figure 10*). If you forget to press this Button, you cannot create the database in the **final step**, because the program will report **"No database selected"** (*cf. figure 18*).

3.2 User Management: Multiple Coders and Permissions

This second step of preparing your DNA workspace allows you to generate multiple user identities with divergent sets of rights for different coders. Thus, you can specify for each coder, which parts of the dataset each user can see or edit and thereby pre-structure your coding and research process. In order to do so, click on second tab **"Coder"** in the sidebar of the "Create new database"-menu (*cf. figure 11*).

In the main window (*cf. figure 11*) you can now see a list with all coders and how many of the 12 possible actions they are permitted to perform. Now you can either add a new user-profile by clicking the **"Add"**-button (*cf. figure 11*) or select an existing coder and adjust her/his users rights by clicking on the user and then on the **"Edit"**-button (*cf. figure 14*). Both options will open the pop-up menu shown in (*cf. figure 12*).

This pop-up menu allows you to configure an individual profile for each coder in three simple steps:

1. You can choose the **colour** for the coder (*cf. 1, figure 12*). It is recommended to choose different—if possible—divergent colours for each coder, because this permits you to detect at the first glance, which user coded which statement, as every coded statement is marked in the

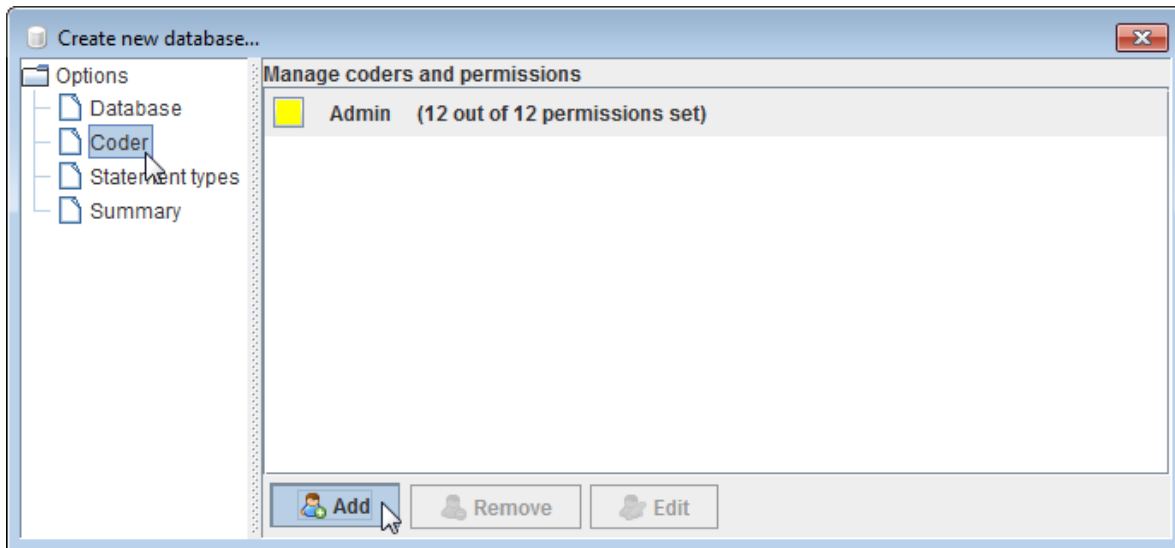


Figure 11: Adding a second coder to the database

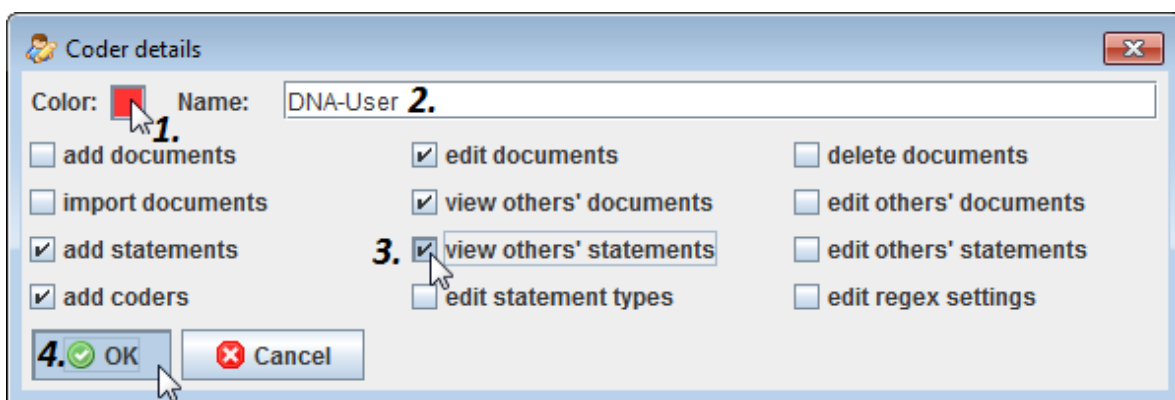
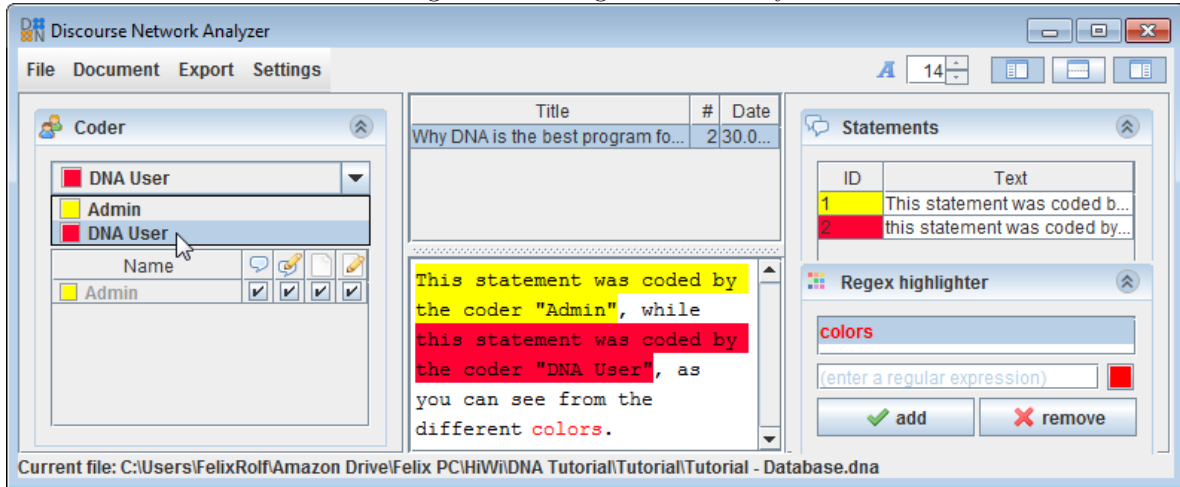


Figure 12: Configuring coder permissions

Figure 13: Change coder identity



individual colour of its respective coder (*cf. middle column of figure 13*).

2. You can enter the preferred name of each coder in the field "**Name**". If possible with respect to data protection rules, it is recommended to use the real names of the coders. This makes it easier for them to select their profile (in the upper left of the main program window) the first time they start the program (*cf. figure 13*).
3. The final step allows you to configure the **permissions** of each coder individually by (de)selecting the respective rights via a click (*cf. 3, figure 12*). Each new user has all of the 12 configurable permissions in the preset mode. Which parts of the dataset an individual coder should be able to see or edit, should depend on your coding process. For better orientation a few practical implications of the 12 configurable permissions are listed in *table 1*. Please keep in mind, that every user can see and change to other user identities either accidentally or because of non-compliance, as s/he has to select her/his role the first time s/he starts the program and can change her/his role anytime (*cf. above & figure 13*)

Finally you approve your choices by clicking the **OK**-button (*cf. 4, figure 12*). It is possible to change the settings either in the "new database"-menu by selecting the respective user and clicking the **Edit**-Button (*cf. figure 13*) or changing the coder settings in the main menu.

Table 1: User permissions explained

Permission	Practical Implication
add documents	The user can add new documents (= raw data) manually (via copy & paste or retyping) to the database ⇒ user has (also) a research function
import documents	The user can import new documents from other sources like .txt or other .dna files to the database or recode the metadata of multiple documents ⇒ user has (also) a research function
delete documents	The user can delete documents from the database or dataset. This option requires at least the other permission "view others' documents" if the user has an organizing or editing function (structuring database for coding by other users) or the permission "add documents" and "add statements" if the coder determines own codes and organizes her/his own set of data.

Table 1: User permissions explained

Permission	Practical Implication
edit documents	The user can edit her/his own documents (= raw data) but not necessarily the codings in these documents that were made by other users—which would require the permission "edit others' statements"—or the documents uploaded by other users—which requires the permission "edit others' documents". This option requires at least the other permission "add documents" or "import documents" and should be selected if the user determines own codes and organizes her/his own set of data or acts as a researcher for the other coders.
view others' documents	The user can view the documents uploaded by other users. This option is necessary for a collaborative coding process in which only a part of the users selects and uploads the raw data (=documents) for all other users. The option should not be selected if each coder comes up with own codes and organizes her/his own set of data.
edit others' documents	The user can edit the documents uploaded by other users. This option requires at least the other permission "view others' documents" and should be selected if a user organizes or edits the raw data provided by other users.
add statements	The coder actually codes the data by creating and editing statements. If only a part of the users select and upload the raw data this option requires the additional permission "view others' documents". If the coder suggests own codes and organizes her/his own set of data this option requires either the additional permission " <i>add documents</i> " or " <i>import documents</i> ".
view others' statements	The coder can view the statements coded by other users. For example the Coder "DNA User" would not see the yellow statement of the Coder "Admin" in figure 13 if this option was deselected for her/his user role. This option should not be selected if you want to establish a blind coding process.
edit others' statements	The coder can edit or correct the statements coded by other users. This option requires at least the other permission "view others' statements" and should only be selected for few users with an <i>organizing controlling or editing function</i> .
add coders	The user can add new coders (<i>see section 3.2</i>). This option should only be selected for few users with an <i>organizing</i> function.
edit statement types	The user can change or complement the variables of interest (<i>see section 3.3</i>). This option should only be selected for very few users or the researchers themselves because possible adjustment of these variables is usually only necessary in cases when the research design and/or research questions change fundamentally.

Table 1: User permissions explained

Permission	Practical Implication
edit regex settings	The user can specify keywords which are highlighted in the text , along with a text color . For example, in <i>figure 13</i> the word "colors" is highlighted in the raw data text (middle column), because it was specified as a keyword in the <i>regex highlighter sidebar</i> in the bottom left of the DNA window. If a user does not have the right to edit the regex setting, the buttons "Add" and "Remove" in this highlighter would be hidden, but the keyword would nevertheless be visibly highlighted in the text and listed in the regex highlighter sidebar. Thus, if you specify a distinct set of theory-based keywords in advance in order to render the coding procedure semi-automatic , you should not enable this option or select it only for <i>few users</i> , as the respective coder could change the keywords. However, if you don't have a theoretically relevant set of keywords in advance or just specify them as a assistance for your coders, you can allow them to formulate such keywords by themselves.

3.3 Statement Types and Variables

Clicking on the third tab in the sidebar of the "Create new database"-menu - "**Statement Types**" (*cf. figure 15*) - opens a menu, which allows you to adjust or supplement either the variables or the types of statements, which your coders derive from the raw data.

3.3.1 Adjusting the variables of interest

The statement type "**DNA Statement**" represents a text portion of your raw data, where an actor reveals her/his opinion/belief/etc. about an issue. Thus, the main task of your coder(s) is to identify such text portions and gain the relevant data about the actor or his opinion/belief/etc. Your research question or theory should not only dictate what kind of information should be coded as statements but also which relevant variables of this information should be captured by the coder. As you can see in the "Statement Types"-menu, DNAs default configuration allows capturing four variables. Selecting "**DNA Statement**" and clicking on the button "**Edit**" (*cf. figure 15*) opens a pop-up window (*cf. figure 16*), which reveals the nature of this four preconfigured variables, along whose lines the coders can collect information:

- the **person** who makes the statement.
- the **organization** the speaker is affiliated with.
- the **concept** (opinion/belief/etc.) which is raised by the actor.
- a dummy variable indicating whether the actor **agrees** with the concept or not.

Furthermore the pop-up window depicted in *figure 16* shows, that each variable is assigned to a specific data type: While "person", "organization" and "concept"—according to their nature as nominal variables—will be coded by a short text, "agreement" as a dichotomous variable will be coded as a **boolean data type** , which accordingly only allows for two forms (either agreement or non-agreement). Neither the data type nor the name of the variables can be changed directly. However by selecting a variable and clicking on the **trash symbol** (on the right side of the "Add Variable"-button, *cf. 4, figure 16*) you can delete a variable and subsequently replace it by a new one. Generating a new

Figure 14: Edit coder details

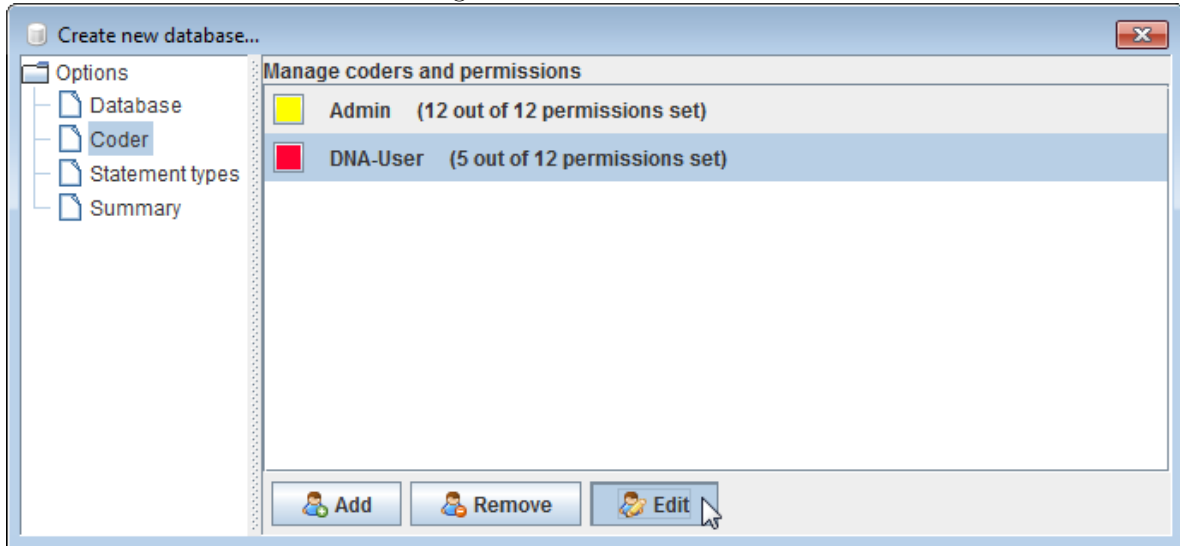
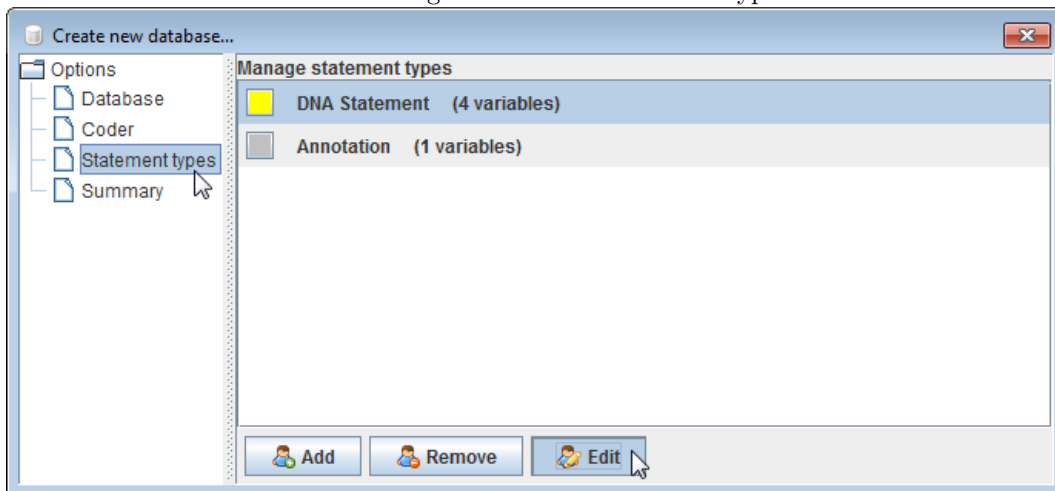


Figure 15: Edit Statement Types



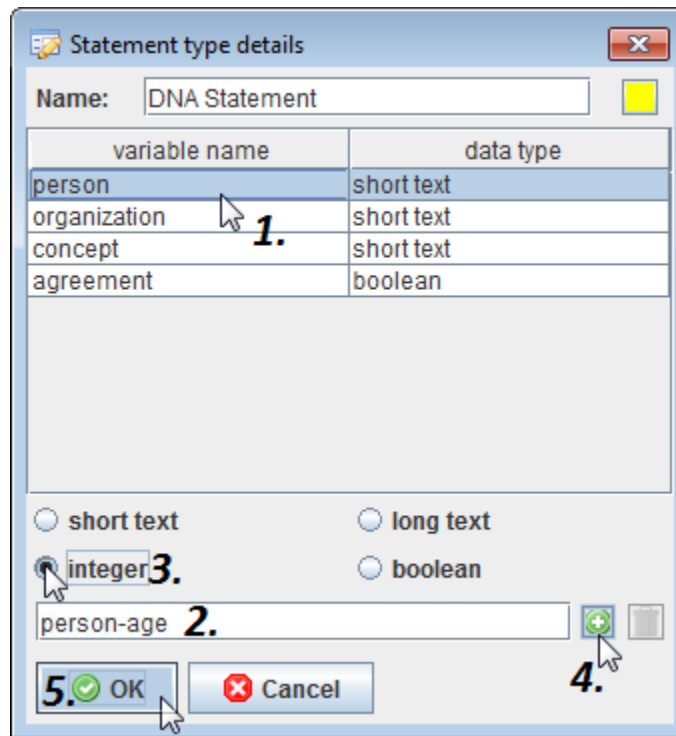


Figure 16: Edit Statement Type details

variable—either to replace one of the preconfigured variables or because you are interested in further or other than the preset variables—is possible in five simple steps:

1. You have to **select a existing variable** in order to activate the variable menu (*cf. 1, figure 16*)
2. Now you can enter the **name** of the new variable in the **text field** at the bottom of the pop-up window (*cf. 2, figure 16*). For example, in *figure 16* we are interested in collecting the age of the person who makes the statement. Please note, that DNA does not allow spaces in variable names: Putting a space in the variable name will disable the "Add Variable"-button necessary for step 4.
3. Now you can choose the **data type** of your variable by **clicking** on one of the **four options**. In our example, we choose the option "integer", as the age of a person is neither a nominal nor a dichotomous variable, but an **integer number** (*cf. 3, figure 16*).
4. You have to click on the **"Add-Variable"**-button, which has the form of a **green plus-symbol** (*cf. 4, figure 16*). If this button is disabled, you probably did not select a existing variable (step 1) or have a space in your variable name (*cf. step 4*).
5. Click the **"OK"**-button to confirm your choices (*cf. 5, figure 16*).

Please note, that—for the statement type "DNA Statement"—you should only specify variables, in which you have an actual research interest and that accordingly have to be coded for all statements by all coders. If you are interested in additional and optional information about some statements, you can specify them as variables of the other preconfigured statement type—**"Annotation"**.

3.3.2 Adjusting the statement types

There are very few research scenarios, in which it is necessary to complement the two existing types of statements with further statement type or the an adjustment of "DNA statement". One of them would be, if you study two parallel yet divergent research questions, which employ the same dataset *and* the same coders at the same time. In this case, you could **first rename** the statement type "DNA Statement" by **selecting** it from the statement type menu, clicking the **"Edit"**-button (*cf. figure reffig:statttype*), entering the new name (in this case: "Statement for Research Project 1") in the **text field on top of the pop-up window** (*cf. figure 16*) and pressing the **"OK"**-button (*cf. 5, figure 16*). Subsequently you would open a **new pop-up window** by clicking on the **"Add"**-Button in the statement type menu (*left button in figure reffig:statttype*), **name** the new statement type (in this case: "Statement for Research Project 2") in the text field on top of the pop-up window and **choose a color** (different from the other type) by clicking on the colored button next to this text field as well as **specify the relevant variables** synchronous to the procedure depicted in section 3.3.1 However, please evaluate carefully, if it is really necessary for your second research interest that you specify a second statement type or if it would be possible to either conceptualize it as a variable of the existing statement type or study it sequentially or with a different set of coders (and therefore in a different DNAdatasets). **More than two statement types (besides "Statement" and "Annotation") can cause a confusion of the coders and therefore compromise the validity of the coding procedure.**

3.4 Final step: Approving your workspace and creating the DNA file

Finally, clicking on the last tab in the sidebar of the "Create new database"-menu - **"Summary"** provides you with a summary of your choices in respect to the configuration of your coding process (*cf. figure 17*). After controlling each of the three information you can now create your database by clicking on the **"Create database"**-button. If this button is disabled and you get the report "'No database selected'" (*cf. figure 18*), you probably forgot to click the **Apply-Button** after specifying your database (*cf. step 3 in section sec:createlocal*). After creating the database, the new database will open in the main DNA window (*cf. figure 5*) and you can proceed towards loading up and organizing the raw data.

4 Using DNA: Importing and Organizing your Raw Data

This section describes how to upload and organize your research project's raw data—i.e. the text files (newspaper articles, press releases etc.) containing the uncoded statements—in DNA.

4.1 Opening an existing DNA database

First of all, you have to choose, in which DNA-Database you want to upload and process your data. Thus, you either **open** an existing DNA-database. To open a DNA-database, simply follow the steps depicted in Figure 1: First, click on the index tab **"File"** and select the option **"Open DNA-database"** (*cf. 1, figure 19*). As a result, a pop-up window will appear, which allows you to choose between opening a **"Local .dna file"** or a **"remote database on a server"**. If your database is stored on an remote server, you should choose the second option and repeat the procedure outlined in **Creating and using a remote database (MySQL)**. If your dataset is stored in a folder on your local PC or device, you can proceed with the preset option and click on the button **"Browse"** (*cf. 2, figure 19*), which will open a further pop-up window, in which you can find your database by choosing its storage location

Figure 17: Summary of your about to be created DNA database

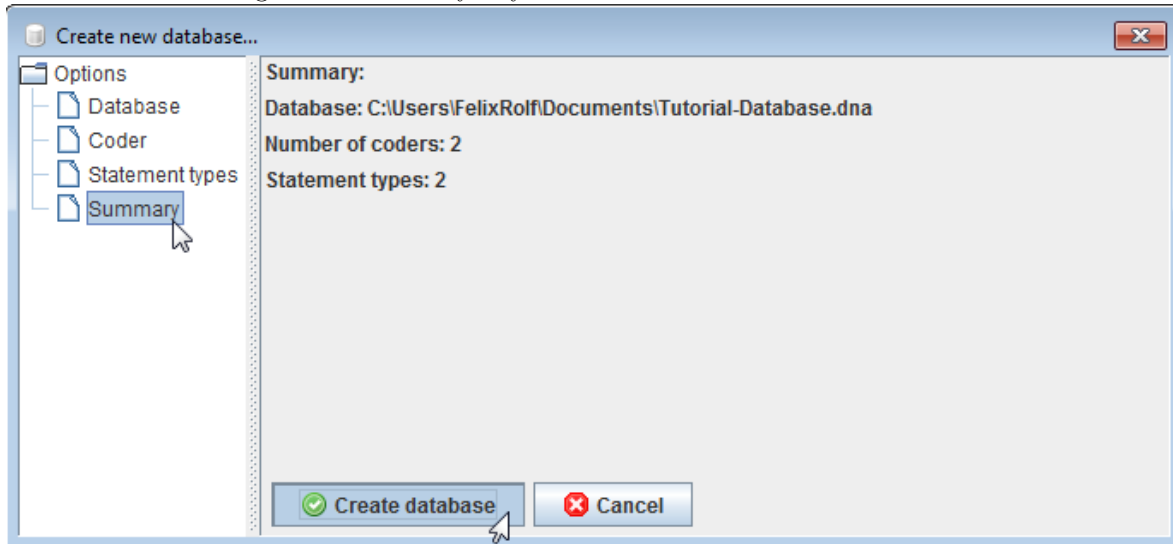


Figure 18: No database selected (e.g. if choice was not applied)

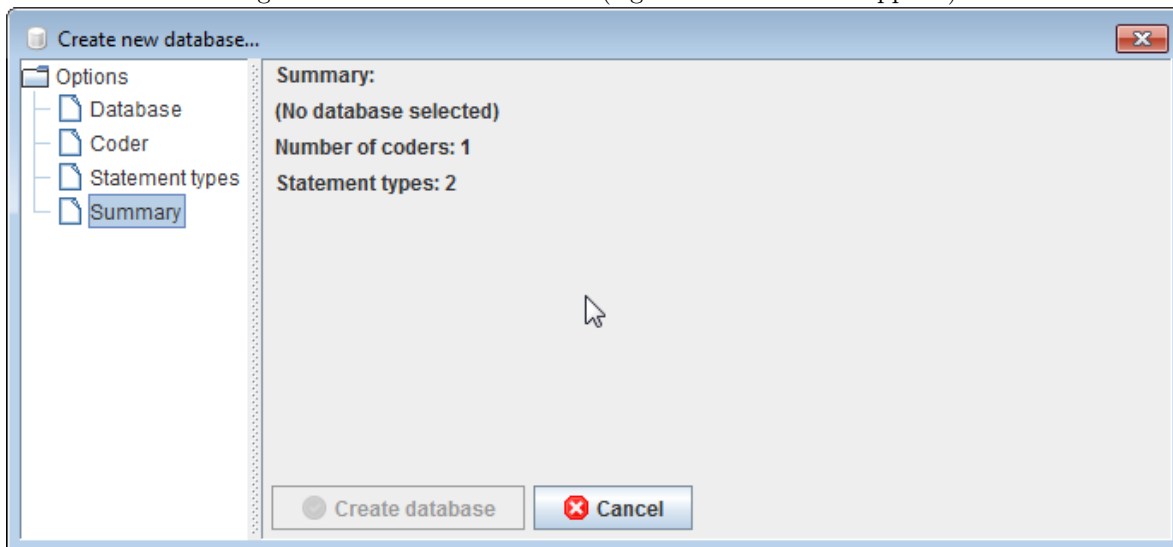
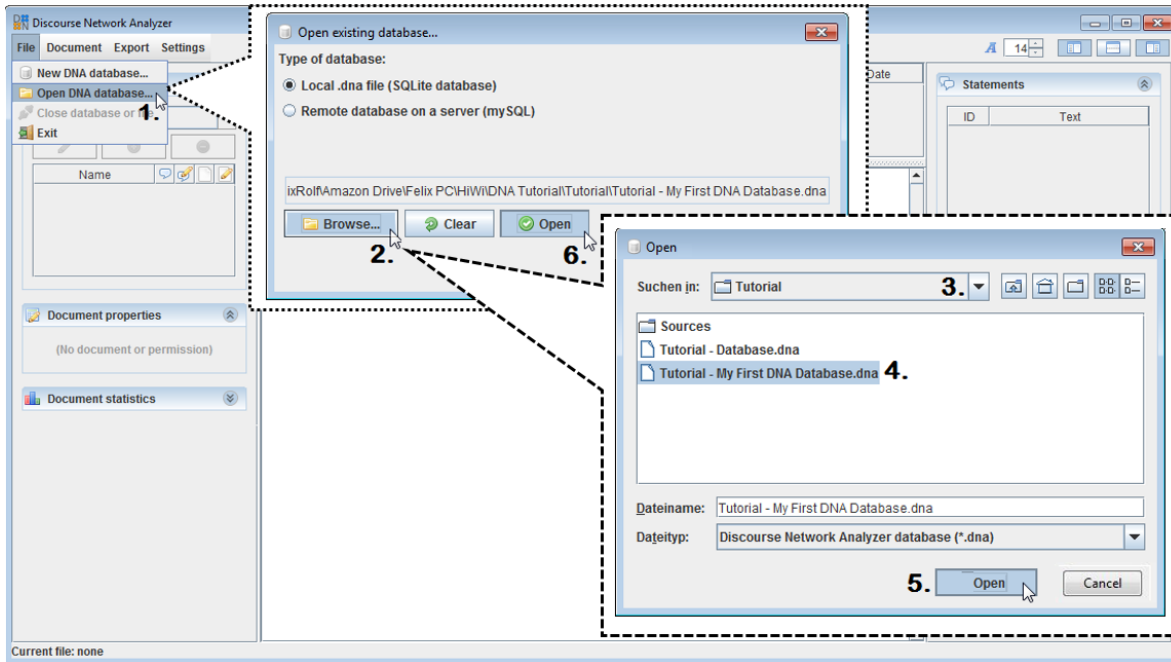


Figure 19: Opene DNA-database



from the "Save in"-slide down menu (*cf. 3*), selecting the respective database (*cf. 4*) and clicking on the button "Open" both in the pop-up- and the "Open existing database..."-window (*cf. 5 & 6*).

4.2 Importing Documents (Raw Data)

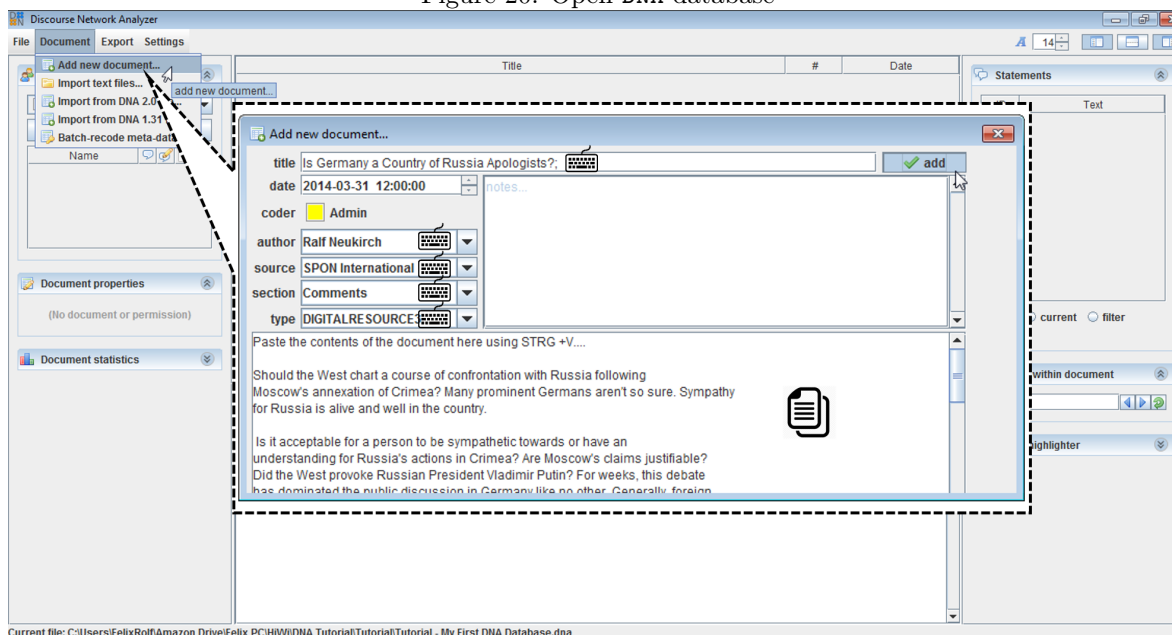
There are four different—partly semi-automatic—ways to upload your raw data and related descriptive information (title, date, author, source, section and type of document) into DNA: Importing single Documents manually via Copy & Paste, Importing multiple Documents semi-automatically from text files, Importing Documents from other DNA databases and using `rDNA` to import data which is already available in R (WIP!). All four will be explained in detail in this section.

4.2.1 Importing single Documents manually via Copy & Paste

The most basic way to import data to DNA requires you to manually copy & paste the content and the descriptive information for each of your documents into the text fields of a pop-up window, which you open by clicking on the index tab "Documents" and selecting the option "Add new document" (*cf. figure 20*). This window has eight text boxes, in which you can enter information from and about your source data (*cf. figure 20*):

- The field "title" is mandatory and may include any kind of information, for instance a unique ID if you plan to collect additional information about the articles in a separate database. Duplicate article titles are not allowed.
- The field "date" is also mandatory and preset on the current time and day. You can change it by either clicking on the year, month, day or time and adjusting the respective value via the arrows on the right or by manually entering the date in the format "YYYY-MM-DD hh:mm:ss".

Figure 20: Open DNA-database



Please make sure you enter the date correctly because otherwise the algorithms for longitudinal data will not work properly.

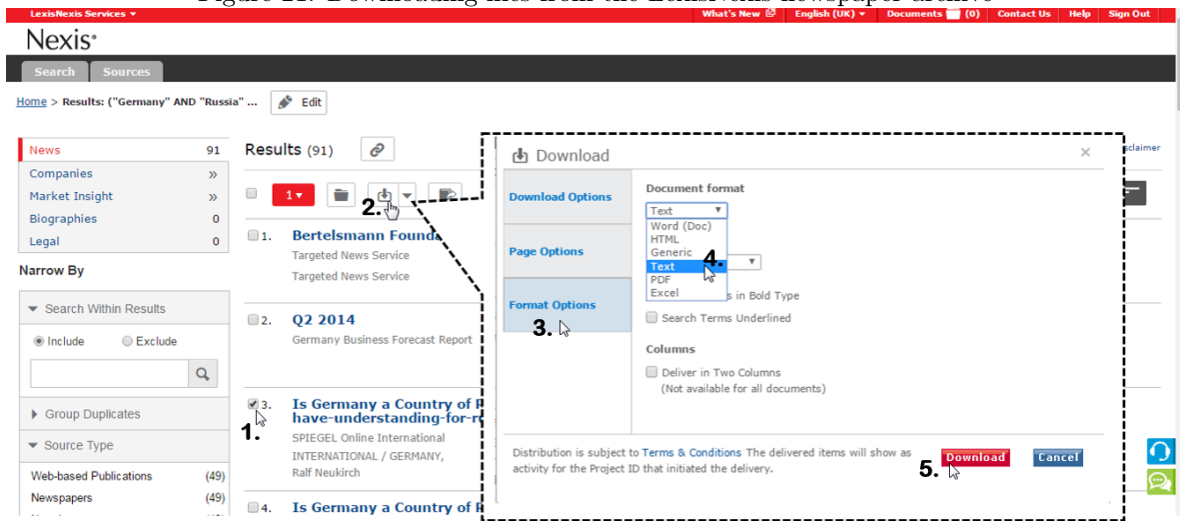
- The fields "author", "source", "section" and "type" are optional, but this additional information can help you to efficiently organize your data and ensure the reproducibility, transparency and future usage of your research project. You can enter this information either manually or select an author, source, section or type you specified for a previously added document from the drop-down menu, which appears when you click of the downward arrow on the left of the respective field.
- To insert the content of your document, copy your article from a website or any other text source and paste it in the **text field (largest field at the bottom of the pop-up-window)**. Single line breaks are automatically removed, while double line breaks (paragraph breaks) are preserved. Some escape sequences and special characters are automatically removed when text is inserted.
- If you want to add further meta-information to your document, which does not fit the preset categories, you can use the field "notes".

Finally—after checking your specifications—you can import the document to DNA by clicking the "Add"-Button.

4.2.2 Importing multiple Documents semi-automatically from text files

However if you want to analyze a greater number of articles, it quickly becomes tedious to manually copy & paste each document and its meta-data. This is why DNA also offers a semi-automatic way to upload multiple documents and their relevant meta-data (author, date, source, type) at the same time.

Figure 21: Downloading files from the LexisNexis newspaper archive



Downloading and Preparing your Raw Data. This way of importing raw data to DNA requires that you save all documents as **separate ".txt"-files** (one file for each article) **in a common folder**. Please note, that you have to use the ".txt"-format for saving your data, as DNA can not import ".doc" or ".pdf"-files. In case you use the newspaper database of LexisNexis—which is available through many university libraries—for finding and retrieving your raw data, please make sure that you download all documents separately (by selecting the individual document before clicking the download button, *cf. step 1-2, figure 21*) and choose the document format "Text" (under "Format Options" in the Download pop-up menu, *cf. step 3-4, figure 21*) before downloading the data (*cf. step 5, figure 21*).⁶

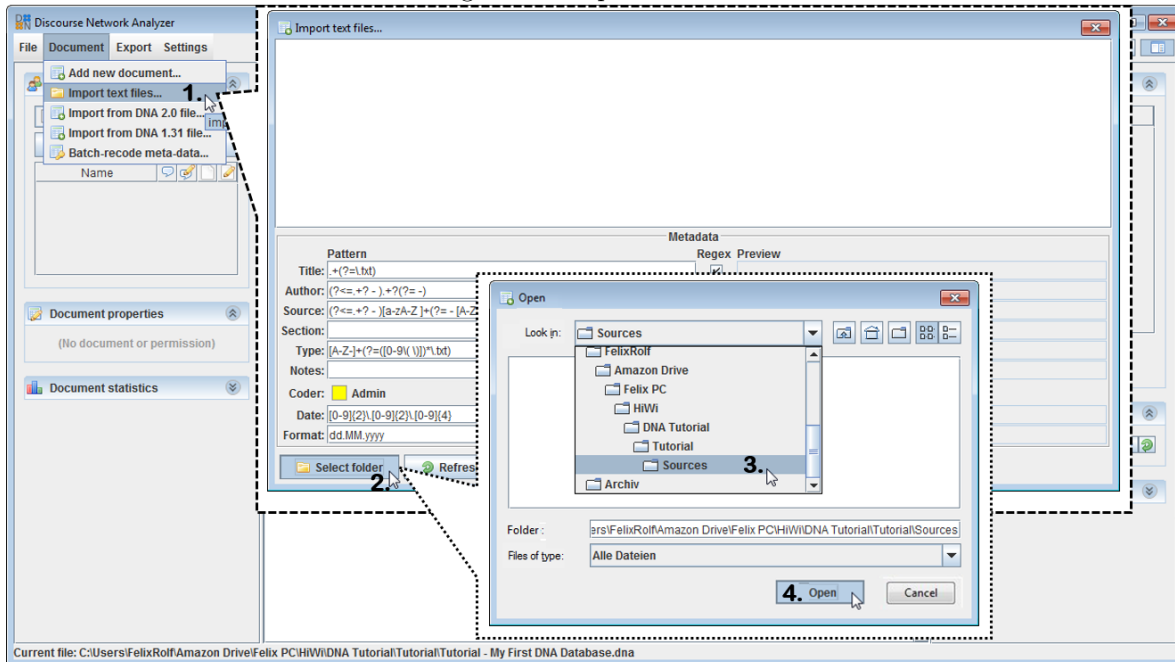
If you want to use the preset regex-configurations (**in contrast to adjusting them**) for automatically detecting and uploading the meta-data of your documents, you should use a **file name** in the format **"DD.MM.YYYY - Author - Source - TYPE.txt"** with blanks before and after the minuses, where "DD.MM.YYYY" is the date, on which the article was published. While "Author" and "Source" do not require a special format or length (e.g. you can use the first and/or last name of the author), the type of the document must always be indicated by capital letters. For example, the file name of the article [spon.de/aecID](#), which is used as an example here, would have the format "31.03.2014 - Ralf Neukirch - SPON International - DIGITALRESOURCE.txt". Please note, that plain text files are sometimes saved as ".TXT" instead of ".txt" files. While this is technically the same, it can cause problems while importing multiple text files. If this is the case, you have to either change the preset Regex-configuration or correct the ".txt"-suffix manually in the file name(s). Otherwise the automatic detection of your documents' meta-data won't work.

Importing your Raw Data into DNA If you prepared your data adequately, you can retrieve the documents and the relevant additional information in four simple steps (*cf. figure 22*):

1. Click on the index tab **"Documents"** and select the option **"Import text files"** (*cf. step 1, figure 22*). As a result, a new window will open, in which you press the button **"Select folder"** (*cf. step 2*). This will open a further pop-up menu. Here, you have to select the **folder**, in which

⁶If you use rDNA it will soon also be possible to import LexisNexis-data into DNA via using rDNA and a new R-package called [LexisNexisTools](#).

Figure 22: Import text files



you saved the text-files of your raw data, from the **"Look in"-slide down menu** (cf. step 3) and click the button **"Open"** (cf. step 4).

2. Now all documents, which are stored in the respective folder, should be listed in the main window of the "Import text files..."-pop up (cf. figure 23). If this isn't the case, please check if your documents are saved in the right file format (.txt). In order to check, whether DNA is able to automatically identify your documents' meta-data, select one of the documents and click on the **"Refresh"**-Button (cf. figure 23). If you specified the file names correctly, you can now see the respective meta-data of the selected document in the fields "Title", "Author", "Source", "Type" and "Date" of the **"Preview"**-Section at the bottom right of the "Import text files"-window (cf. figure 23).
3. If you want to adjust or amend the meta-data manually, just select the document, **uncheck** the box **"Regex"** of the field you want to edit and enter the divergent/additional information in the field on the left. Then click again on the **"Refresh"**-Button to check, whether you changes were accepted.
4. Finally, click on the button **"Import files"** to import all documents of the respective folder into your DNA database (you do not need to select each document for import).

Adjusting the Regex-Configuration for automatic identification of meta-data. The previous steps assumed that you use the preset configuration of DNA to detect and upload the meta-data (Title, Author, Source, Type, Date) of your documents automatically into your database. However, if you are interested in automatically importing additional information about your source data (in the fields **"Section"** or **"Notes"**) or if your file names naming system layed out here (but nevertheless contain all relevant information in a systematic order), DNA allows you to change, adjust or amend the pattern, through which the meta-data about your documents is derived from the file names. The

Figure 23: Import text files

Pattern	Regex	Preview
Title: .+(?=\.txt)	<input checked="" type="checkbox"/>	31.03.2014 - Ralf Neukirch - SPON International - DIGITALRESOURCE
Author: (?<=.+? -).+?(?=-)	<input checked="" type="checkbox"/>	Ralf Neukirch
Source: (?<=.+? -)([a-zA-Z]+(?= - [A-Z0-9\ \])+\.txt)	<input checked="" type="checkbox"/>	SPON International
Section:	<input checked="" type="checkbox"/>	
Type: [A-Z]+(?=((0-9\ \]))*\.txt)	<input checked="" type="checkbox"/>	DIGITALRESOURCE
Notes:	<input checked="" type="checkbox"/>	
Coder: Admin		
Date: [0-9]{2}\.[0-9]{2}\.[0-9]{4}	<input checked="" type="checkbox"/>	31.03.2014
Format: dd.MM.yyyy		Mon Mar 31 00:00:00 CEST 2014

Buttons: Select folder Refresh Import files

commands/rules, on which the "translation" of file names into meta-data is based, are formulated in the **Regular expressions (in short: Regex)-Syntax** and can be edited for each kind of information (Title, Author, Source, Section, Tyoe, Notes, Date) in the field **"Pattern"** on the bottom left of the "Import text files..."-window (cf. figure 23). If you want to amend or adjust this settings it is recommended to use a Regex Cheatsheet (see e.g. cheatography.com or [this regex "translator"](#)). As further support, figure 24 translates the preset regular expressions of the DNA-"Import text files..."-option.

4.2.3 Importing Documents from other DNAdatabases

You can also import documents from other DNAdatabases. This function is particularly relevant in two scenarios: First, if you not only want to use the **raw data but also the coded statements** of an already finished research project, this function allows you to import both. Secondly, if there is **more than one person working on the same project at the same time** and you did not use multiple user roles (see section 3.2) to enable your coders to work on the same remote database. In the second scenario, you should use this function to prepare your datasets or merge the codings, as it is usually difficult to merge the files manually later on. In the latter scenario, the function helps you to avoid trouble with **duplicate statement IDs** and article names, as DNA will take care of e.g. duplicates automatically.

Make sure, that you **know which version of DNA** (DNA 2.0 or older) was used to create and edit the database, from which you want to import data, **before** using the "Import from DNA"-function. If you use this manual as a beginner's Tutorial for working with DNA please download the

file "sample.dna" from the DNA github.com/leifeld/dna/releases. This file contains a small selection of documents and statements from a larger project about congressional hearings on climate change [Fisher et al. \(2013a,b\)](#).

To import documents (and the included code statements), click on the index tab **"Documents"** and

Figure 24: Import text files

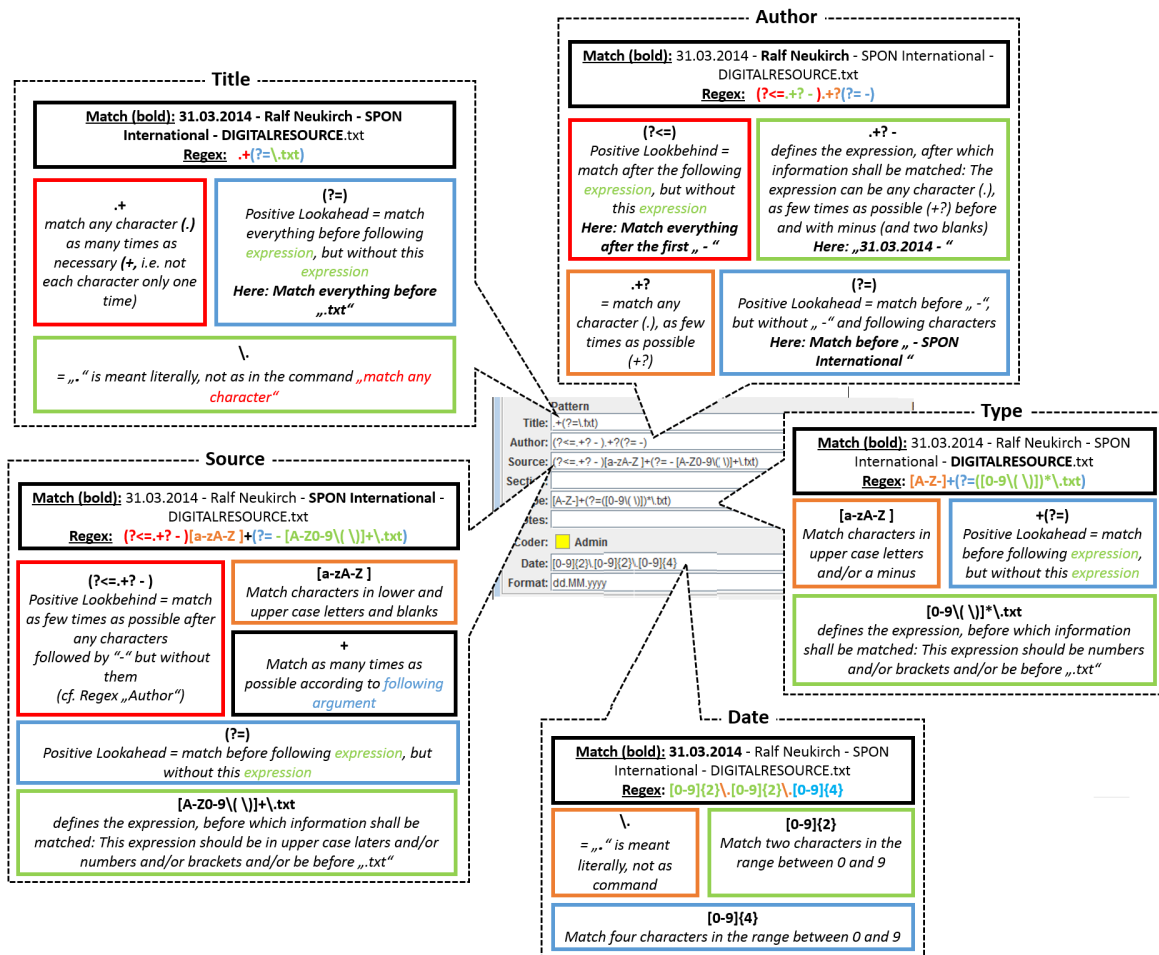
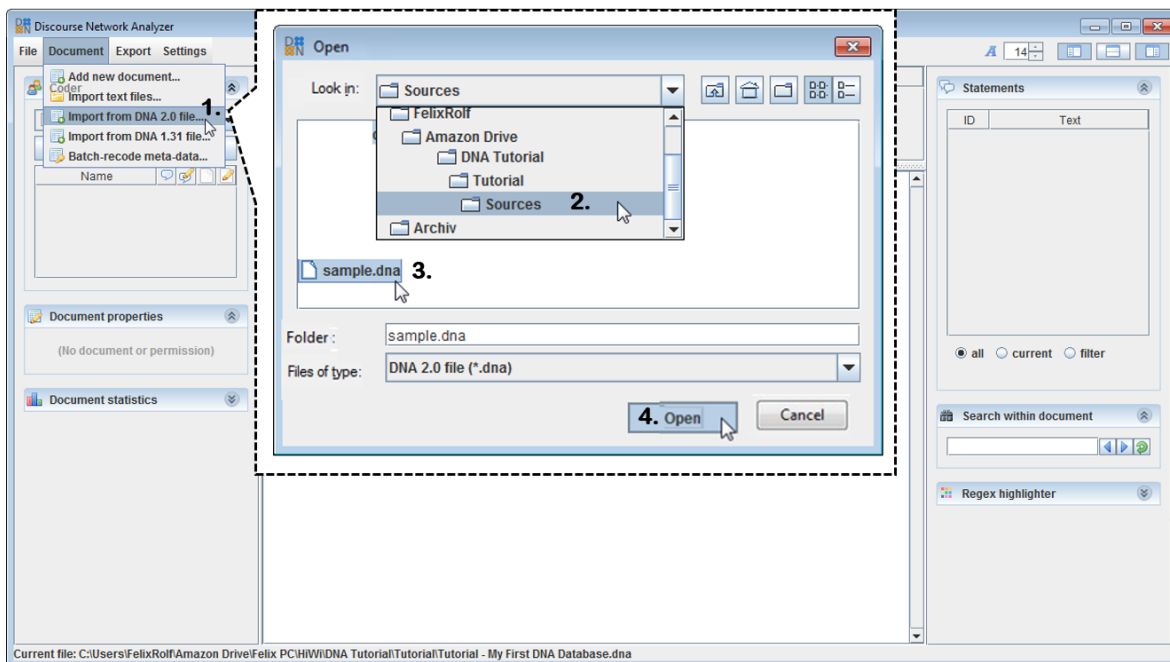


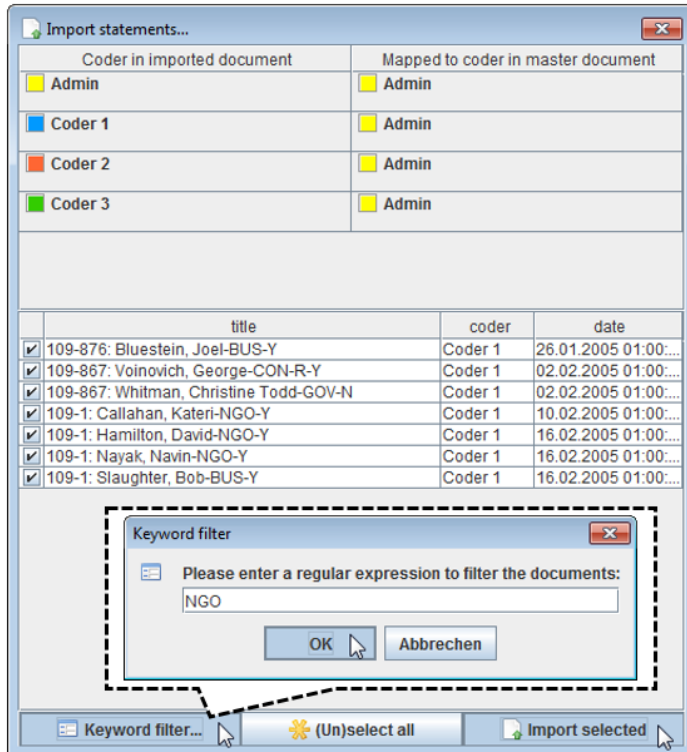
Figure 25: Import a DNA 2.0-database



select the option **"Import from DNA 2.0 file"**, if DNA 2.0 or was used to create and edit the database. As the file type has of .dna files has significantly changed since version 1.31, databases created with and older version of DNA need to be impored using the seperate method **"Import from DNA 1.31 file"** (cf. step 1, figure 25). As a result, a further pop-up menu will open (cf. figure 26). In this window, you have to select the **folder**, in which you saved the text-files of your raw data, from the **"Look in"-slide down menu** (cf. step 2) and **select the respective .dna file** (cf. step 3). Click the button **"Open"** (cf. step 4) will then open the menu depicted in figure 26.

In this menu, you can select, which documents (and respective which coded statements) from the original DNAdatabase you want to import in your database by either manually checking or unchecking the boxes on the left of the document title or by using the function **"Keyword filter"**. This function is particularly helpful if you want to only import few documents with a specific common characteristic (author, topic) from a very large dataset. Clicking on the button **"Keyword filter..."** (cf. left button in figure 26) opens a new pop-up window, in which you can enter a specific search term. For example, if you downloaded and opened the **"sample.dna"**-file, you can select all congressional hearings of NGO representatives by **entering the keyword "NGO"** in the text field and pressing the button **"OK"** in the "Keyword filer" pop-up window (cf. figure 26). Now only the boxes of the three documents, which contain the hearings of NGO representatives Kateri Callahan, David Hamilton and Nayak Navin, should be checked, while the other boxes are unchecked. The "Keyword filter.."-Function is based on the same regex syntax described in **Adjusting the Regex-Configuration for automatic identification of meta-data**. This means, you can also use more specified regular expressions (cf. figure 24 or *regex cheatsheet*) to select certain articles. For example, if you enter a **"^N"** in the "Keyword filter"-text field DNA will select all articles starting with a capital N. If you want to undo your selections, you can also automatically select or unselect all articles by pressing the button **"(Un)select all"** in the middle of the "Import statements"-window (cf. figure 26). Pressing the right button **"Import selected"** in the same window imports all documents with a checked box (and the respective coded statements) in your DNAdatabase (cf. figure 26). If you use this manual as a beginner's tutorial for working with DNA,

Figure 26: Import Statments menu



you should try importing all documents and the respective statements from the file "sample.dna" into your database.

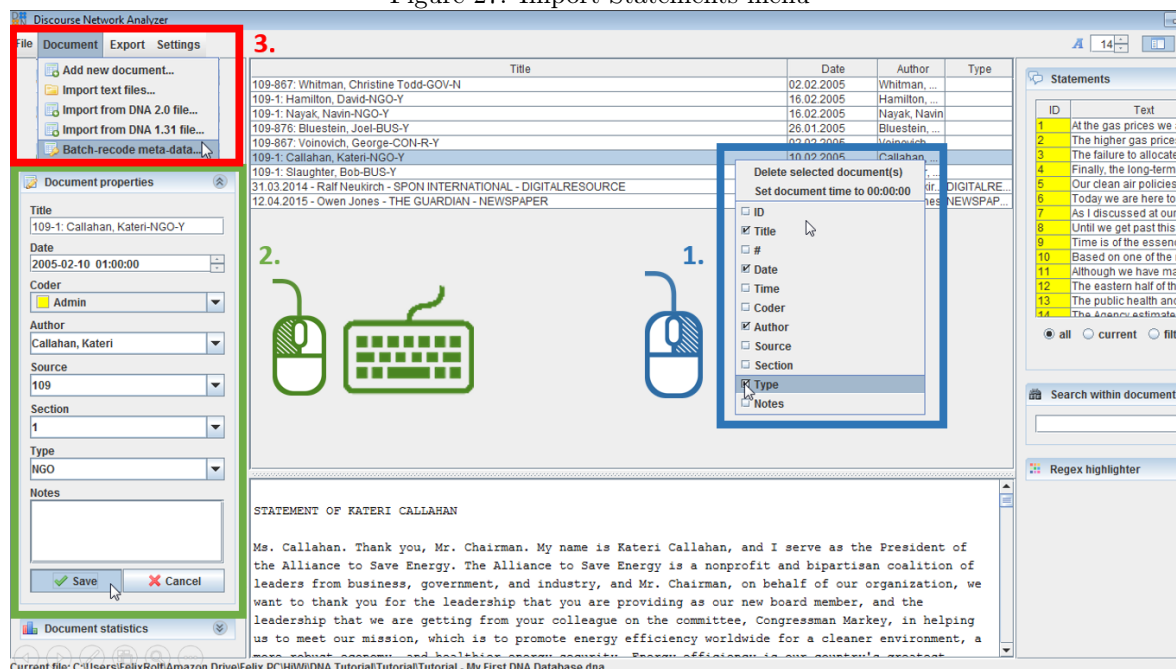
4.3 Organizing documents (Raw Data)

4.3.1 Deleting and navigating through documents

All your imported documents are listed in the upper middle table of theDNA main window. If you click on an article, its corresponding text (i. e. the speech) will be displayed in the text area below the document table. By clicking on, for example, the entry "109-1: Callahan, Kateri-NGO-Y" you open the speech of Kateri Callahan, a representative of the Alliance to Save Energy. You can not only adjust the size of the document table (by clicking on the bar above the text area and moving it vertically with your cursor) or its columns (by clicking on the edge of the column and moving it horizontally with your cursor), but also customize the meta-information, which are displayed in the document table: Just **right click on any document** and use the appearing **context-menu** to (un-)check the boxes of the information you (don't) want to be displayed (*cf. figure 27, step 1*). A structured (and customised) overview of your raw-data is essential for detecting missing information and thus efficiently controlling, organizing and coding your data. For example, if you display the meta-information "Type" (by checking the respective box in the context-menu), you can see that the type of all documents from the sample.dna file is not listed.

The same context menu can be used to delete documents from your database by **selecting the documents** you want to delete (pressing and holding the "**Ctrl**"-key for selecting multiple documents), opening the context menu with a **right click** and choosing the option "**Delete selected documents**".

Figure 27: Import Statements menu



4.3.2 Editing the documents' meta-data (author, time etc...)

DNA allows you to edit, delete or complement the descriptive information related to your raw data (title, date, author, source, section and type of document). Similiar to the procedures outlined in section 4.2 there is a manual as well as a semi-automatic way to adjust the meta-data of your documents.

Editing the documents' meta-data manually. The most basic way to edit your documents' meta-data is to **select the document**, of which you want to edit the information (by left-clicking on it) and adjusting the values in the **"Document properties"**-submenu on the middle left of the DNA main window (cf. figure 27, step 2) by either manually typing in the relevant information or by selecting an already specified author, a source, a section or a type from the **drop-down menu** on the right of the respective meta-field. For example, in figure 27 (step 2) Kateri Callahans speech was selected, and the value "NGO" (for Non-Governmental Organisation) was manually specified as "Type of document" by entering it in the field "Type" of the "Document properties"-submenu. Do not forget to **press the button "Save"** in the submenu (cf. figure 27, step 2) to confirm your edits.

Please note, that you can manually only edit the meta-data of **one document at one time**. If you try to select multiple documents for editing, the "Document properties"-submenu will disappear, returning "(No document or permission)"

Editing the documents' meta-data semi-automatically. However if you want to adjust the meta-data of a greater number of articles, it quickly becomes tedious to manually edit information about each document. This is why DNA also offers a semi-automatic way to edit, delete or complement the descriptive information related to your documents. In order to edit your documents' meta-data semi-automatically, click on the index tab **"Documents"** and select the option **"Batch-recode meta-data"** (cf. step 3, figure 27). As a result, a pop-up window similar to figure 28 10 will open. In the

Figure 28: Meta information recode window

ID	Source field	Old target field	New target field
1	109-876: Bluestein, Joel-BUS-Y		
2	109-867: Voinovich, George-CON-R-Y		
3	109-867: Whitman, Christine Todd-GOV-N		
4	109-1: Callahan, Kateri-NGO-Y		NGO
5	109-1: Hamilton, David-NGO-Y		NGO
6	109-1: Nayak, Navin-NGO-Y		NGO
7	109-1: Slaughter, Bob-BUS-Y		
9	31.03.2014 - Ralf Neukirch - SPON INTERNATIONAL - DIGITALRESOURCE	DIGITALRESOURCE	DIGITALRESOURCE
8	12.04.2015 - Owen Jones - THE GUARDIAN - NEWSPAPER	NEWSPAPER	NEWSPAPER

upper half of this pop-up window you find nine fields, which can be configured in order to adjust the meta-data for **multiple documents at once**:

- The field **"Target field:"** specifies, which kind of meta-information (i.e. title, author, source, section, type, notes) should be adjusted by choosing the respective meta-data category from the slide- down menu (which you open by clicking the arrow on the right of the target field).
- The field **"Source field:"** specifies, where the data you want to use for adjusting the target field is stored. For example, if you simply want to delete or correct (e.g. misspelled) title-, author-, source-, section-, type- or notes-metadata, you usually choose the same field as source field as you have chosen as target field, since you want to adjust the data already stored in this field. However, if you want to add new data to a (maybe empty or incomplete) target field, you have to choose the part of the meta-information as source field, which contains the information, from which you want to derive the new data. As the document title should contain all relevant meta-information, "Title" is usually used as source field for the latter case.
- The field **"Matching on target regex"** allows you to automatically delimit the documents which you want to adjust, based on the information stored in the document's target field. Similiar to all regex-implementations in DNA you can either use search terms or regular expressions to filter the documents. If you, for instance, misspelled the author "Ralf Neukirch" sometimes as "Ralf Neunkirch", you can correct all your misspellings by simply selecting "Author" as "Target field", entering "Ralf Neunkirch" in the field "Matching on target regex:" and the correct version ("Ralf Neukirch") in the field "New target field". As "Matching on target regex" automatically deselects all non-matching cases (here: All documents, who do not have "Ralf Neunkirch" specified as their author), the meta-information (here: "Author") remains the same for all other documents.
- The field **"Matching on source regex"** similarly allows you to automatically filter the documents of which you want to alter the meta-data, based on the information stored in the document's source field. For example, if you realise that Ralf Neukirch does not write for "SPON

International" (as you erroneously specified), but for "THE GUARDIAN", you can simply correct all your misspecifications by first selecting "Source" as the "Target field" and "Author" as the "Source field", secondly entering "Ralf Neukirch" in the field "Matching on source regex" and then specifying "THE GUARDIAN" as "New target field".

- The field **"%target regular expression"** allows you to specify/match a part of the target field, which you want to use as new information in the same field. For example, if the field "Author" somehow contains the full document titles you can reduce the information in the field "Author" to just the name of the respective author by entering the regular expression "(?<=.+?—).+?(?= -)" (cf. figure 24 or regex cheatsheet) in the field "%target regular expression" and entering "%target" in the field "New target field". **Please note, that if you do not use this function, you should not change the preset value ".+" in this field**—because if you do, your recoding might not obtain the expected results.
- The field **"%target replacement"**—similarly to the fields "New target field" and "%source replacement"—defines a new value for the information in the target field. If you use "%target" as "New target field", you have to specify the new/additional/corrected/reduced information in this field.
- The field **"%source regular expression"** allows you to specify/match a part of the source field, which you want to use as new information in the target field. For example, if your source field is "Title" and the titles of your documents have the recommended format (i.e. "DD.MM.YYYY - Author - Source - TYPE.txt" with blanks before and after the minuses; see section 4.2.2) you can automatically specify the meta-information for the field "Author" by (1) choosing "Author" as the "Target field" and "Title" as the "Source field", (2) entering the regular expression "(?<=.+?—).+?(?= -)" (cf. figure 24 or regex cheatsheet) in the field "%source regular expression" and (3) entering "%source" in the field "New target field". **Please note, that if you do not use this function, you should not change the preset value ".+" in this field**—because if you do, your recoding might not obtain the expected results.
- The field **"%source replacement"**—similarly to the fields "New target field" and "%target replacement"—defines a new value for the information in the target field. If you use "%source" as "New target field", you have to specify the new/additional/corrected/reduced information in this field.
- The field **"New target field"** defines the new/corrected/reduced/additional data, which is entered in your target field (cf. examples above). Please note, that this field has to be set on "%source" (preset value) if you use the functions "source regular expression" or "source replacement" and has to be set on "%target" if you use the functions "target regular expression" or "target replacement". Otherwise, the respective functions will not work.

The lower half of the "Recode document meta-data"-pop-up window (cf. figure 28) displays a table with four columns and a row for each of your documents, which help you to preview, control and trace back your changes to the meta-data:

- The column **"ID"** contains the individual ID of each of your documents. This column can be particularly helpful if you specify a recoding procedures for a certain set of documents. If you know the ID of a few exemplary documents from this set, you can quickly trace back and understand the consequences of your recoding-specifications by scrolling down to the respective ID and taking a look at the other columns of these documents.
- The column **"Source field"** displays the field, from which you get the meta-data for recoding the target field. It is particularly helpful to understand the sequence of information in the source field, if you want to specify a "%source regular expression" or use "Matching on source regex" (for example, if only some source fields contain the relevant information).

- The column **"Old target field"** shows the meta-data in the target fields prior to your adjustments. It is particularly helpful if you want to use "%target regular expression" or use "Matching on target regex" (for example, if you only want to change the value of a certain set of target fields)
- The column **"New target field"** displays the consequences of your adjustment. It is particularly helpful to control for the successfulness of your recoding and undesired recodings (for example, if the target field already contained the relevant information but is recoded nevertheless).

Your recodings are only applied, if you press the button **"Recode"** (on the lower right of the "Recode document meta-data"-window, cf. figure 28). **Once this is applied, it cannot be undone!** So please control the consequences of your recodings by using the table at the lower half of the window. However, before pressing the "Recode"-button, you *can* revert all adjustments by pressing the button **"Revert changes"** and therefore are able to experiment with the meta-data (regex) specifications.

As noted previously, all documents from the file "sample.dna" do not specify any meta-data concerning the type of the respective document. Both figure 28 and figure 29 illustrate an exemplary semi-automatic procedure for complementing this information based on the information stored in the document-title (here: The organisation, to which the respective speaker belongs to). Thus in both examples, "Type" is selected as "Target field", while "Title" is selected as "Source field".

The example in figure 28 uses **manual search terms** to specify the meta-information for the document-type. By entering "NGO" in the field "Matching on source regex" the adjustments are limited to the documents, which contain "NGO" in the document title. By entering "NGO" in the field "New target field", the new value for "Type" is specified for the selected documents. As you can see in the table on the lower half of the "Recode meta-data"-window, this very simple procedure is insofar successful, as only the target-fields of documents containing hearings of NGO-representatives are changed and the target fields of all other documents (including those with already correct "Type"-information) remain unchanged. However, this procedure would have to be repeated for each kind of organisation from the sample (NGO, GOV, BUS).

The more elegant and less elaborate way of semi-automatically specifying meta-information is depicted in figure 29, which uses the **Regex-syntax**. Here, by entering "^?" in the field "Matching on target regex", only those documents are selected for amendment, which do not already contain any information about the document-type (therefore excluding those documents with already correct "Type"-information). By specifying "(?<=.{0,1})[A-Z]+" as "%source regular expression" (and accordingly "%source" as "New target field"), DNA is instructed to filter any string of upper-case characters before a minus in the document title and set it as a new value for "Type". Thus you can recode the document-type for all documents at once, ensuring that already specified values are not overwritten—as evident from the table in the lower half of the window.

5 Using DNA: Coding the Data

6 Using DNA: Exporting the coded Data

Figure 29: Meta information recode window (regex explained)

The screenshot shows the 'Recode document meta-data' window. It has fields for 'Target field' (set to 'Type'), 'Source field' (set to 'Title'), and various regex options. A table at the bottom shows the mapping of source fields to new target fields. Callouts on the right explain specific regex parts: '^\$' for matching empty fields, '(?<=.+?-)[A-Z]+' for positive lookbehind and uppercase letters, and '.+?-' for matching characters before and after a specific expression.

Regex: $^{\$}$
Match only empty fields (here: fields, which do not have any value for „Type“)

Regex: $(?<=.+?-)[A-Z]^+$
 $(?<=)$
Positive Lookbehind = match after the following *expression*, but without this *expression*

$.+?-$
defines the expression, after which information shall be matched: The expression can be any character ($.$), as few times as possible ($+?$) before and with minus

$[A-Z]$
Match only characters in upper case letters

$^+$
match the characters as many times as necessary ($^+$, i.e. not only one character)

ID	Source field	Old target field	New target field
1	109-876: Bluestein, Joel-BUS-Y		BUS
2	109-867: Voinovich, George-CON-R-Y		CON
3	109-867: Whitman, Christine Todd-GOV-N		GOV
4	109-1: Callahan, Kateri-NGO-Y		NGO
5	109-1: Hamilton, David-NGO-Y		NGO
6	109-1: Nayak, Navin-NGO-Y		NGO
7	109-1: Slaughter, Bob-BUS-Y		BUS
9	31.03.2014 - Ralf Neukirch - SPON INTERNATIONAL - DIGITALRESOURCE	DIGITALRESOURCE	DIGITALRESOURCE
8	12.04.2015 - Owen Jones - THE GUARDIAN - NEWSPAPER	NEWSPAPER	NEWSPAPER

7 rDNA: Using DNA from R

DNA can be connected to the statistical computing environment R (R Core Team 2014) through the rDNA package (Leifeld 2017). There are two advantages to working with R on DNA data.

The first advantage is replicability. The network export function of DNA has many options. Remembering what options were used in an analysis can be difficult. If the analysis is executed in R, commands—rather than mouse clicks—are used to extract networks or attributes from DNA. These commands are saved in an R script file. This increases replicability because the script can be re-used many times. For example, after discovering a wrong code somewhere in the DNA database, it is sufficient to fix this problem in the DNA file and then re-run the R script instead of manually setting all the options again. This reduces the probability of making errors and increases replicability.

The second advantage is the immense flexibility of R in terms of statistical modelling. Analysing DNA data in R permits many forms of data analysis beyond simple visualization of the resulting networks. Examples include cluster analysis or community detection, scaling and application of data reduction techniques, centrality analysis, and even statistical modelling of network data. R is also flexible in terms of combining and matching the data from DNA with other data sources.

7.1 Getting started with rDNA

The first step is to install R. Installing additional R packages for network analysis and clustering, such as `statnet` (Goodreau et al. 2008; Handcock et al. 2008, 2016), `xergm` (Leifeld et al. 2017a,b), `igraph` (Csardi and Nepusz 2006), and `cluster` (Maechler et al. 2017), is recommended. Moreover, it is necessary to install the `rJava` package (Urbanek 2016), on which the rDNA package depends, and the `devtools` package (Wickham and Chang 2016), which permits installing R packages from GitHub (see section 2.4).

```
install.packages("statnet")
install.packages("xergm")
install.packages("igraph")
install.packages("cluster")
install.packages("rJava")
install.packages("devtools")
```

After installing these supplementary packages, the **rDNA** package can be installed from [GitHub](#). The **devtools** package contains a function that permits easy installation of R packages from [GitHub](#) and can be used as follows to install **rDNA**:

```
library("devtools")
install_github("leifeld/dna/rDNA")
```

Once installed, the **rDNA** package must be attached to the workspace:

```
library("rDNA")
```

To ensure that the following results can be reproduced exactly, we should set the random seed in R:

```
set.seed(12345)
```

Now we are able to use the package. The first step is to initialize DNA. Out of the box, **rDNA** does not know where the DNA **.jar** file is located. We need to register DNA with **rDNA** to use them together. To do that, one needs to save the DNA **.jar** file to the working directory of the current R session and then initialize DNA as follows (with **dna-2.0-beta19.jar** in this example):

```
dna_init("dna-2.0-beta20.jar")
```

After initializing DNA, we can open the DNA graphical user interface from the R command line:

```
dna_gui()
```

Alternatively, we can provide the file name of a local DNA database as an argument, and the database will be opened in DNA. For example, we could open the **sample.dna** database that is provided for download on [GitHub](#) under Releases:

```
dna_gui("sample.dna")
```

For this to work, the database file has to be saved in the working directory of the R session, or the path needs to be provided along with the file name.

In addition to opening the GUI, we will want to retrieve networks and attributes from DNA. For this to happen, a connection with a DNA database must first be established using the **dna_connection** function:

```
conn <- dna_connection("sample.dna")
```

The `dna_connection` function accepts a file name of the database including full or relative path (or, alternatively, a connection string to a remote MySQL database) and optionally the login and password for the database (in case a remote MySQL database is used). Details about the connection can be printed by calling the resulting object called `conn`.

After initializing DNA and establishing a connection to a database, we can now retrieve data from DNA. We will start with a simple example of a two-mode network from the sample database. To compute the network matrix, the connection that we just established must be supplied to the `dna_network` function:

```
nw <- dna_network(conn)
```

The resulting matrix can be plotted using visualization functions from the `statnet` suite of packages:

```
library("statnet")
gplot(nw)
```

It is also easily possible to retrieve the attributes of a variable, for example the colours and types of organizations, using the `dna_attributes` function:

```
at <- dna_attributes(conn)
```

The result is a data frame with organizations in the rows and one column per organizational attribute. The next section will provide usage examples of both the `dna_network` and the `dna_attributes` functions.

7.2 Retrieving networks and attributes

This section will explore the `dna_network` function and facilities for retrieving attributes in more detail. The `dna_network` function has a number of arguments, which resemble the export options in the DNA export window. The help page for the `dna_network` function provides details on these arguments. It can be opened using the command

```
help("dna_network")
```

We will start with a simple example: a one-mode congruence network of organizations in a policy debate. The `sample.dna` database is a small excerpt from a larger empirical research project that tries to map the ideological debates around American climate politics in the U.S. Congress over time. Details about the dataset from which this excerpt is taken are provided by [Fisher et al. \(2013a,b\)](#). Here, it suffices to say that the `sample.dna` file contains speeches from hearings in the U.S. Congress in which interest groups and legislators make statements about their views on climate politics. Accordingly, one should expect to find a polarized debate with environmental groups on one side and industrial interest groups on the other side. To compute a one-mode congruence network, the following code can be used:

```
congruence <- dna_network(conn,
  networkType = "onemode",
  statementType = "DNA Statement",
  variable1 = "organization",
  variable2 = "concept",
  qualifier = "agreement",
  qualifierAggregation = "congruence",
  duplicates = "document")
```

The result is an organization \times organization matrix, where the cells represent on how many concepts any two actors (i.e., the row organization and the column organization) had the same issue stance (by values of the qualifier variable `agreement`).

The arguments of the `dna_network` function resemble the options in the DNA export window. Details on the various arguments of the function can be obtained by displaying the help page (`?dna_network`).

`statementType = "DNA Statement"` indicates which statement type should be used for the network export. In this case, the statement type `DNA Statement` contains the variables `organization`, `concept`, and `agreement`. The argument `qualifierAggregation = "congruence"` causes `rdna` to count how often the unique elements of `variable1` have an identical value on the `qualifier` variable (here: `agreement`) when they refer to a concept (`variable2`).

If the algorithm finds duplicate statements within documents—i.e., statements containing the same organization, concept, and agreement pattern—, only one of them is retained for the analysis (`duplicates = "document"`).

The resulting matrix can be converted to a network object and plotted as follows:

```
nw <- network(congruence)
plot(nw,
      edge.lwd = congruence^2,
      displaylabels = TRUE,
      label.cex = 0.5,
      usearrows = FALSE,
      edge.col = "gray"
    )
```

Here, we used the `edge.lwd` argument of the `plot.network` function to make the line width proportional to the strength of congruence between actors. We used squared edge weights to emphasize the difference between low and high edge weights. We also displayed the labels of the nodes at half the normal size, suppressed arrow heads, and changed the colour of the edges to grey. More information about the visualization capabilities of the `network` and `sna` packages are provided by Butts (2008a,b, 2015).

The network is not particularly polarized. We can suspect that some of the concepts are not very contested. If they are supported by all actors, this may mask the extent of polarization with regard to the other concepts. From our experience with the dataset, we can tell in this particular case that the concept “There should be legislation to regulate emissions.” is in fact very consensual. If everybody agrees to this concept, it obfuscates the real structure of the network. Therefore we should exclude it from the congruence network. To do that, we need to export and plot the congruence network again and use the `excludeValues` argument this time:

```
congruence <- dna_network(conn,
                          networkType = "onemode",
                          statementType = "DNA Statement",
                          variable1 = "organization",
                          variable2 = "concept",
                          qualifier = "agreement",
                          qualifierAggregation = "congruence",
                          duplicates = "document",
                          excludeValues = list("concept" =
                                                "There should be legislation to regulate emissions.))
nw <- network(congruence)
plot(nw,
```



```

edge.lwd = congruence^2,
displaylabels = TRUE,
label.cex = 0.5,
usearrows = FALSE,
edge.col = "gray"
)

```

This reveals the structure of the actor congruence network. There are two camps revolving around environmental groups on the right and industrial interest groups and state actors on the left, with **Energy and Environmental Analysis, Inc.** taking a bridging position. The strongest belief congruence ties can be found within, rather than between, the coalitions.

Next, we should tweak the congruence network further by changing the appearance of the nodes. We can use the colours for the organization types saved in the database and apply them to the nodes in the network. We can also make the size of each node proportional to its activity. The `dna_attributes` function serves to retrieve these additional data from DNA. The result is a data frame with the relevant data for each organization in the `colour` and `frequency` columns:

```

at <- dna_attributes(conn,
                     statementType = "DNA Statement",
                     variable = "organization")
at

```

##	id	value	color	type	alias	note
## 1	16	Alliance to Save Energy	#00CC00	NGO		
## 2	7	Energy and Environmental Analysis, Inc.	#FF9900	Business		
## 3	14	Environmental Protection Agency	#000000	Government		
## 4	25	National Petrochemical & Refiners Association	#FF9900	Business		
## 5	11	Senate	#000000	Government		
## 6	19	Sierra Club	#00CC00	NGO		
## 7	22	U.S. Public Interest Research Group	#00CC00	NGO		
##	frequency	in dataset	in network			
## 1	4	TRUE	TRUE			
## 2	4	TRUE	TRUE			
## 3	11	TRUE	TRUE			
## 4	4	TRUE	TRUE			
## 5	5	TRUE	TRUE			
## 6	6	TRUE	TRUE			
## 7	6	TRUE	TRUE			

To use these data in the congruence network visualization, we can use the plotting facilities of the `plot.network` function:

```

plot(nw,
     edge.lwd = congruence^2,
     displaylabels = TRUE,
     label.cex = 0.5,
     usearrows = FALSE,
     edge.col = "gray",
     vertex.col = at$color,
     vertex.cex = at$frequency
)

```


This yields a clear visualization of the actor congruence network, with simultaneous display of the network structure including its coalitions, the actors' activity in the debate, and actor types.

Another way to visualize a discourse network is a two-mode network visualization. To compute a two-mode network of organizations and concepts, the following code can be used:

```
affil <- dna_network(conn,
  networkType = "twomode",
  statementType = "DNA Statement",
  variable1 = "organization",
  variable2 = "concept",
  qualifier = "agreement",
  qualifierAggregation = "combine",
  duplicates = "document")
```

This creates a 7×6 matrix of organizations and their relations to concepts. The argument `networkType = "twomode"` is necessary because the rows and columns of the `affil` matrix should contain different variables. The arguments `variable1 = "organization"` and `variable2 = "concept"` define which variables should be used for the rows and columns, respectively. The arguments `qualifier = "agreement"` and `qualifierAggregation = "combine"` define how the cells of the matrix should be populated: `agreement` is a binary variable, and the `combine` option causes a cell to have a value of 0 if the organization never refers to the concept, 1 if the organization refers to the respective concept exclusively in a positive way, 2 if the organization refers to the concept exclusively in a negative way, and 3 if there are both positive and negative statements by the organization about the concept. `rDNA` reports on the R console what each combination means.

As in the previous example, the resulting network matrix can be converted to a `network` object (as defined in the `network` package). The colours of the edges can be stored as an edge attribute, and the resulting object can be plotted with different colours representing positive, negative, and ambivalent mentions.

```
nw <- network(affil, bipartite = TRUE)
colors <- as.character(t(affil))
colors[colors == "3"] <- "blue"
colors[colors == "2"] <- "red"
colors[colors == "1"] <- "green"
colors <- colors[colors != "0"]
set.edge.attribute(nw, "color", colors)
plot(nw,
  edge.col = get.edge.attribute(nw, "color"),
  vertex.col = c(rep("white", nrow(affil)),
    rep("black", ncol(affil))),
  displaylabels = TRUE,
  label.cex = 0.5
)
```

In this example, we first converted the two-mode matrix to a bipartite `network` object, then created a vector of colours for the edges (excluding zeros), and inserted this vector into the `network` object as an edge attribute. It was necessary to work with the transposed `affil` matrix (using the `t` function) because the `set.edge.attribute` function expects edge attributes in a row-wise order while the `as.character` function returns them in a column-wise order based on the `affil` matrix. Finally, we plotted the network object with edge colours and labels. In the visualization, we used white nodes for organizations and black nodes for concepts.

We can now see the opinions of all actors on the various concepts. The blue edge indicates that **Energy and Environmental Analysis, Inc.** has both positive and negative things to say about the concept "Emissions legislation should regulate CO2. This is why this organization acts as a bridge between the two camps in the congruence network. Furthermore, we can now see more clearly that the concept we omitted in the congruence network, "There should be legislation to regulate emissions", is viewed positively by four organizations but still receives a negative mention by one actor. The green edges span both camps, and this caused additional connections between the two groups. The negative tie is ignored in the construction of the congruence network because conflicts are not counted and there is no second red tie to that concept.

References

- Butts, C. T. (2008a). Social network analysis with **sna**. *Journal of Statistical Software*, 24(6):1–51.
- Butts, C. T. (2008b). **network**: A package for managing relational data in R. *Journal of Statistical Software*, 24(2):1–36.
- Butts, C. T. (2015). **network**: *Classes for Relational Data*. The Statnet Project (<http://statnet.org>). R package version 1.13.0.
- Csardi, G. and Nepusz, T. (2006). The **igraph** software package for complex network research. *Inter-Journal, Complex Systems*, 1695(5):1–9.
- Fisher, D. R., Leifeld, P., and Iwaki, Y. (2013a). Mapping the ideological networks of American climate politics. *Climatic Change*, 116(3):523–545.
- Fisher, D. R., Waggle, J., and Leifeld, P. (2013b). Where does political polarization come from? Locating polarization within the U.S. climate change debate. *American Behavioral Scientist*, 57(1):70–92.
- Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., and Morris, M. (2008). A **statnet** tutorial. *Journal of Statistical Software*, 24(9):1–26.
- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Bender-deMoll, S., and Morris, M. (2016). **statnet**: *Software Tools for the Statistical Analysis of Network Data*. The Statnet Project (<http://www.statnet.org>). R package version 2016.9.
- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). **statnet**: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1–11.
- Leifeld, P. (2017). **rdna**: *R bindings for the Discourse Network Analyzer*. R package version 2.0.1.
- Leifeld, P., Cranmer, S. J., and Desmarais, B. A. (2017a). *Temporal Exponential Random Graph Models with btergm: Estimation and Bootstrap Confidence Intervals*. Forthcoming.
- Leifeld, P., Cranmer, S. J., and Desmarais, B. A. (2017b). **xergm**: *Extensions of Exponential Random Graph Models*. R package version 1.8.2.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2017). **cluster**: *Cluster Analysis Basics and Extensions*. R package version 2.0.6.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Urbanek, S. (2016). **rJava**: *Low-Level R to Java Interface*. R package version 0.9-8.
- Wickham, H. and Chang, W. (2016). **devtools**: *Tools to Make Developing R Packages Easier*. R package version 1.12.0.