**Apêndice do artigo Desafiando as Fronteiras do Jornalismo por meio de Objetivos Comunicativos, de Christoph Raetzsch e Martin Brynskov (em inglês)**

**Operations log file for dataset:** 170630_RADENTSCHEID_Media Mentions_List
**Dataset created:** 03 July 2017
**Log created**: 03 July 2017
**Creator**: C. Raetzsch
**Source**: https://volksentscheid-fahrrad.de/de/medienspiegel/

**Introduction:** This file documents the steps taken to convert a website (SOURCE) into a spreadsheet OBJECT for further analysis. We include this TABLE with the ARTICLE to illustrate how this process involves a range of decisions and operations on data to convert between one digital OBJECT (a website) into a new one (spreadsheet). The initial data is a test OBJECT for such routines, which could also be applied to data sources compiled by researchers.

The cleaning and standardizing of data is not as straight forward as is often claimed in the context of digital methods. We also include the observations on data that we made and the conclusions we draw. The aim of documenting this process is to make transparent the research process at a very basic level. The data table is hosted along with the article for further analysis (LINK).

This is an *exploratory study of the digital properties* of the website entries and what research routines can be applied to these without expert knowledge or additional resources beyond a single laptop. The routines documented here can be replicated and illustrate the constitution of an epistemic object, or an object of knowledge, from a mere list of entries on a website (see article).

**GET: Data**

- Website (SOURCE) contains a list of media mentions of „Radentscheid-Initiative“ compiled by members of the initiative. Entries have dates, name of medium, title of article and web links (embedded). We use the structure of the website to extract data for research purposes.

**GET: Data through front end**

- copy/paste entire table to a spreadsheet, using TableTools2 (Firefox plugin) to keep tabulation intact.
- List is more or less accurately formatted. 150+ lines had mistakenly been merged into a single cell. That was corrected. Format of dates has been checked and standardised.
- Results in a list of 923 mentions (with dates, the medium, and headlines). No links. Table Tools2 does not copy link locations, only the front-end text “öffnen” (open link)

**GET: Data through source**

- copy selection source of the website (HTML) into a text editor.

**CLEAN and STANDARDIZE Data**

- Search and replace "<a href="" with a line break. [isolates links] --- applies to 921 cases.
- Search and replace "" target="_blank"" with a line break --- applies to 921 cases.
- **RESULT:** A messy TEXT file, in which the lines starting with http:// or https: are now separated from the rest of the source text.
- Transfer text file to TextWrangler (https://www.barebones.com/products/textwrangler/) to isolate only the links in the order that they were found on the page.
- The isolation of all links (977) takes a number of search and replace routines applied to the entire body of text. The result is a list of 923 mentions of articles and 977 links. The problem is ill-understood at this point. Why are there more links embedded than are listed in the front-end?
- Requires a manual check of links and headlines to find the mistake.

**LESSONS learned**

- numerous links had been placed in wrong cell, some links were twice in the list, which shifted the list down. Manual check is indispensable but at an average of 2 seconds per item, it is feasible. Identify faulty links and manually replace them in the table.
- Referrer information in links is a nuissance: e.g. http://www.berliner-kurier.de/berlin/kiez---stadt/initiative-volksentscheid-kaum-in-fahrt--wird-das-neue-radgesetz-ausgebremst-25586476?originalReferrer=&originalReferrer=https://www.google.de/
- Some referrer information may be useful as links are remembered or copied from Twitter and Whatsapp. Isn't Link literacy a requirement for webmasters?
- Initial copy/paste from source selection missed out on 56 links. Added them by replicating the routine for http:// selection from SOURCE.
- Removed several items that had no link or that did not refer to journalistic content (2 Announcements of events, one article that was no longer available).Double items that were listed under two different dates.
- Added a few links that were missing or not copied from source selection
- Consolidate/Standardize names (spelling of Media outlets): Understand the patterns and the deviations. Sort by name to see deviations of spelling or synonymous meanings

**RESULT:** Standardized List in a TABLE Format

- A list of 919 Media mentions (6 May 2015 to 30 June 2017) and links to respective online content.

**ANALYSE: Data**

- Copy information to new sheet
- Code entries for categories of media: Categories are derived from types of media present in the dataset.

    - Berlin Newspaper/Print+Online (daily)
    - National Newspaper (daily/weekly)
    - Weekly News Magazine (print+online)
    - Berlin City Magazine
    - Non-Berlin Newspaper (German)
    - International Media (non-German-language)
    - Special Interest Publication Bike (print and/or online)
    - Online-only News Site
    - Blog (individual)
    - Activist (Platform or information site)
    - Public Broadcasting (TV and radio)
    - Private Broadcasting-national (TV and radio)
    - Private Broadcasting-local (TV and radio)
    - Berlin City Magazine
    - Non-Berlin Newspaper (German)
    - Miscellaneous

**REFLECTION on Coding**

- Single coder only. In a research setting, train several coders and redo the coding. Code categories are *derived from a dataset*, not the other way around. Usually, a code book is developed beforehand to be applied to a dataset later.
- Decisions on coding certain media need to be reflected. Especially newspapers like *Tagesspiegel* and *taz* are also national news media but were classified as Berlin Newspaper since they are produced in Berlin and cover Berlin themes a lot. Large readership is located in Berlin and the readers of both are also likely to be supportive of the initiative. *WELT* is also focused on Berlin but perceived as a national daily paper. Some blogs are also activists, some online-only media have a blog, etc. Some special interest publications are also online-only sites, some blogs (individual) are also activists. For identifying patterns among the media that are mentioned by the initiative such concerns would change the coding pattern if done in a team.
- Background knowledge that goes into the classification and coding needs to be reflected and documented.

**RESULTS:** New Objects

Based on a table with standardized data formats, new operations are now possible, which could not be applied to the website SOURCE. Transferring information on the SITE to a TABLE is based on an analysis of patterns and deviations in DATA, as much as it requires manual cleaning and standardizing. Because a TABLE OBJECT is often the base for further analysis, we have documented this process in detail. We can now look at patterns within the link collection through the creation of different kinds of objects, that may become useful for analysis and research, depending on the research question. The point here is that in an exploratory research design, many of the OBJECTS can be easily created based on a table and may in fact *generate the questions rather than answering them.*

**CHART**

- Count occurrences of categories in the dataset. COUNTIF-function
- Display category occurrences and count (relative): 919 occurences coded and counted
- Create Pie CHART to show absolute values or percentages.
- **RESULT:** local newspapers are by far the most attentive observers of the initiative. OR, these are the media that activists reference because they read/watch them. Public Broadcasting pays a lot more attention to this issue. OR, is the preferred medium that activists consume. National Newspapers and Online only news sites are prominent media as are weekly news magazines with national reach.
- **REFLECTION:** The point here is not to arrive at a definite conclusion. It seems just as plausible to draw a conclusion about the media preferences of those who run the initiative as it is arguably a finding about who supports or covers such issues the most (local papers and public broadcasters).

**WordCloud**

- Create a wordle from headline text
- Copy entire column of headline to http://wordle.net/ —shows the absolute occurrence of a term relative to the size of the text, simple visualization of dominant patterns
- **RESULT:** Most frequent terms are Berlin/Berliner; Fahrrad [Bike]; Volksentscheid [Referendum]; Senat [Berlin city government]; Radgesetz [bike law]; Initiative
- **REFLECTON:** Obviously, the occurrence of “Berlin” and “Fahrrad” is not surprising, this is the theme and context of the initiative. Terms like “Volksentscheid” [Referendum] are also to be expected, since this is what the initiative is about. Since the initial link collection is made by the initiative itself, it is not surprising either that such coverage is mentioned and linked, which explicitly mentions the “Volksentscheid” [Referendum]. However, based on a collection of sources over a two year period, the relative occurrence of “Volksentscheid” [Referendum] shows that media attention emerges as an issue becomes represented by actors (founding of the initiative) and that has a conflict which is locatable (initiative vs. city government). In the data collection, the articles referring to “Volksentscheid” [Referendum] cover a period of only several months.

**LINK ANALYSIS (not included in paper)**

- Launch DMI Link Ripper (https://wiki.digitalmethods.net/Dmi/ToolLinkRipper) to collect outlinks from all entries in dataset.
- copy entire column LINK to Link Ripper
- Link Ripper provides *all outlinks from every single link* contained in the table.
- **NEW OBJECTS:** a network-file of these outlinks (gdf) and a word/txt file of the outlinks, sorted by host link
- **NOTE**: copy/pasting all LinkRipper output to word file creates a heavy load on the CPU
- Processing finished: A new word file contains 464,000 words or 2,093 pages (all outlinks from each of the 919 links in the TABLE). This volume is too big for bulk operations on a standard laptop.
- Copy/Paste the same LinkRipper Output to a spreadsheet. Now 55,500 lines of links.

- **RESULT:** Different types of linking can be observed among the different media categories. Commercial Media and bloggers tend to insert many internal links.

**Possible Options for further analysis.**

A) Due to heavy server load and volume of output, a reasonable selection of Categories of Medium should be made e.g. compare Activist media / Blogs (individual) and Newspapers from Berlin.

B) Limit the volume to all articles that where published between May and June 2016, when the signature collection took place. Then compare linking to the phase when coalition talks took place (September/October 2016).

This would require more mutations of OBJECTS: more TABLES, NETWORK FILES, IMAGES, TEXT FILES. A process like this can continue ad infinitum …