22_02_2023

Weekly writing progress

- Eligibility criteria
- (Venue selection process)

Upcoming writing progress

- Information sources
- Incl. / Excl. criteria
- Search Strategy

Paper graphs - Takeaways from “Genealogies and Citations ”

Cohesive subgroups (Chapter 3)

We present a number of techniques to detect cohesive subgroups […], all of which are based on the ways in which vertices are interconnected.
Density is the number of lines in a simple network, expressed as a proportion of the maximum possible number of lines.
A **complete network** is a network with maximum density.

The **degree** of a vertex is the number of lines incident with it.
The **indegree** of a vertex is the number of arcs it receives.
The **outdegree** is the number of arcs it sends.
Two vertices are **adjacent** if they are connected by a line.

A **path** is a walk in which no vertex in between the first and last vertex of the walk occurs more than once.
A network is **strongly connected** if each pair of vertices is connected by a path.
A **strong component** is a maximal strongly connected subnetwork.

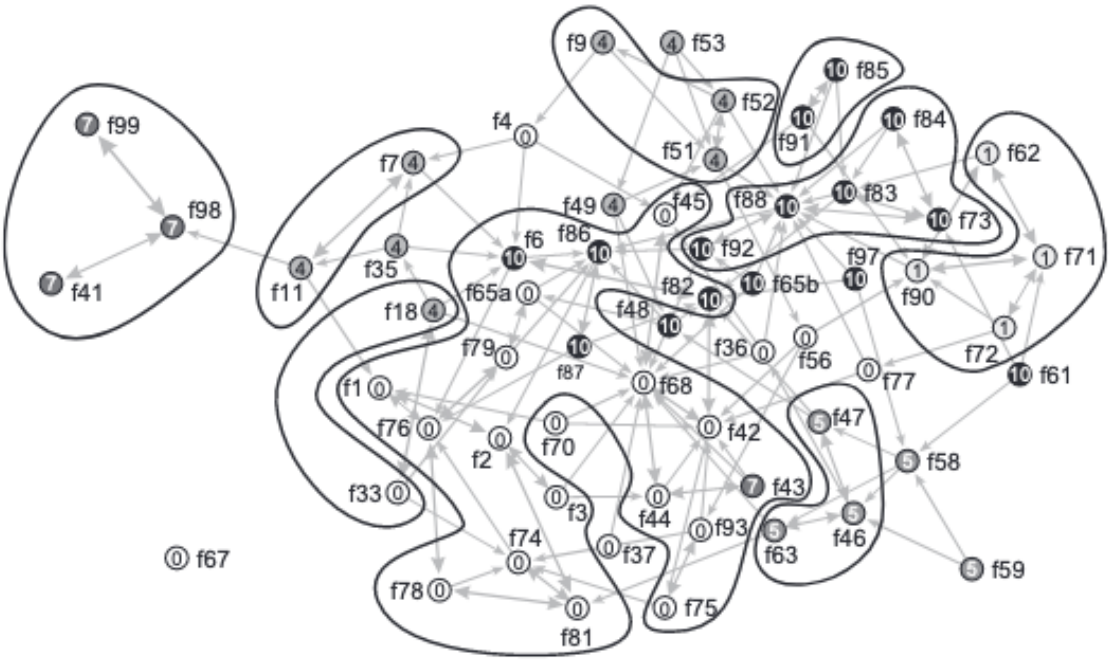


Figure 33. Strong components (contours) and family–friendship groupings (vertex colors and numbers) in the network of Attiro.

k-Cores

The distribution of degree reveals local concentrations of ties around individual vertices, but it does not tell us whether **vertices with a high degree are clustered or scattered all over the network**.

For k-cores, we pay no attention to the degree of one vertex but to the degree of all vertices within a **cluster**.

These clusters are called **k-cores** and k indicates the minimum degree of each vertex within the core

DEFINITION:

"A **k-core is a maximal subnetwork** in which **each vertex has at least degree k** within the subnetwork."

WHY?:

k-cores help to **detect cohesive subgroups**: Remove the lowest k-cores from the network until the network **breaks up into relatively dense components**.

In large networks, this is an effective way of finding cohesive subgroups.

k-cores in literature graphs (Chapter 11)

Researchers operating within a particular subject area or scientific **specialty** tend to cite each other and common precursors.

Citation analysis reveals such cohesive subgroups.

Network analysis is the preferred technique for extracting specialties and research traditions from citations.

Basically, **specialties are cohesive subgroups in the citation network**. In this regard k-cores offer a penetrating view.

Figure 104. *k*-Cores in the centrality literature network (without isolates).

The network contains a 10-core of twenty-nine papers that is the central summit of this network.

Each of the articles in this core is connected to at least ten other articles by citations.

Main path analysis (Chapter 11)

The cohesion concept does not take time into account.

A special technique for citation analysis was developed that explicitly focuses on the flow of **time**. It is called **main path analysis**.

Here, we want to identify the publications that are the **crucial links** in the literature on a particular topic.
Citations indicate how knowledge **flows** through a scientific community.
Each flow follows a path of citations, and **citations that occur in many paths** (i.e. with high weights) are important to the transmission of knowledge.
Citations with high traversal weights are linked into **main paths**, which represent the **main lines of development in a research area**.
The articles and authors connected by citations of some minimum traversal weight constitute main path components, which are hypothesised to identify scientific specialties or subspecialties.

Main path analysis calculates the extent to which a particular citation or article is needed for linking articles.
First, the procedure counts all paths from each **source** (an article that is not citing within the data set) to each **sink** (an article that is not cited within the data set), and it counts the number of paths that include a particular citation.
Next, it divides the number of paths that use a citation by the total number of paths between source and sink vertices in the network. This proportion is the traversal weight of a citation. In a similar way, you can obtain the traversal weight of each article.

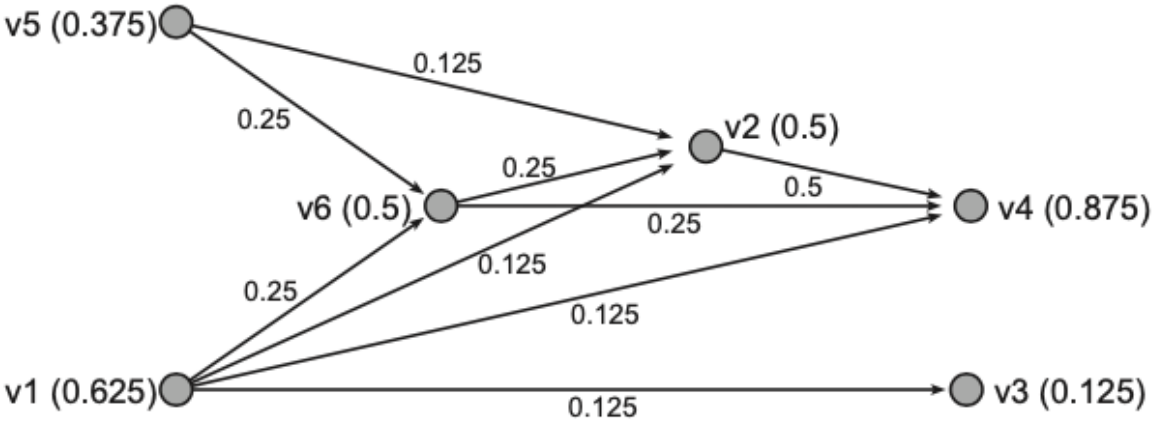


Figure 105. Traversal weights in a citation network.

The example shows a citation network of six articles ordered in time from left to right.
There are two sources (v1 and v5) and two sinks (v3 and v4).
One path connects source v1 and sink v3, but there is no path from v5 to v3.
Four paths reach v4 from v1 and three paths from v5.

In sum, there are eight paths from sources to sinks.
The citation of article v1 by article v3 is included in one of the eight paths, so its traversal weight is 0.125.
The citation of v2 in article v4 is contained in exactly half of all paths.

The traversal weights of the vertices, which are reported between brackets, are calculated in a similar way.

Now that we have defined and calculated the traversal weights of citations, we may **extract** the paths or components with the highest traversal counts on the lines, the **main paths or main path components**, which are hypothesised to identify the **main stream of a literature**.
We can analyse their evolution over time and search for patterns that reflect the integration, fragmentation, or specialisation of a scientific community.

Example of main paths extracted using different techniques:

Figure 106. Forward local and key-route local (top), standard global and key-route global (middle), and backward local (bottom) main paths in the centrality literature network.

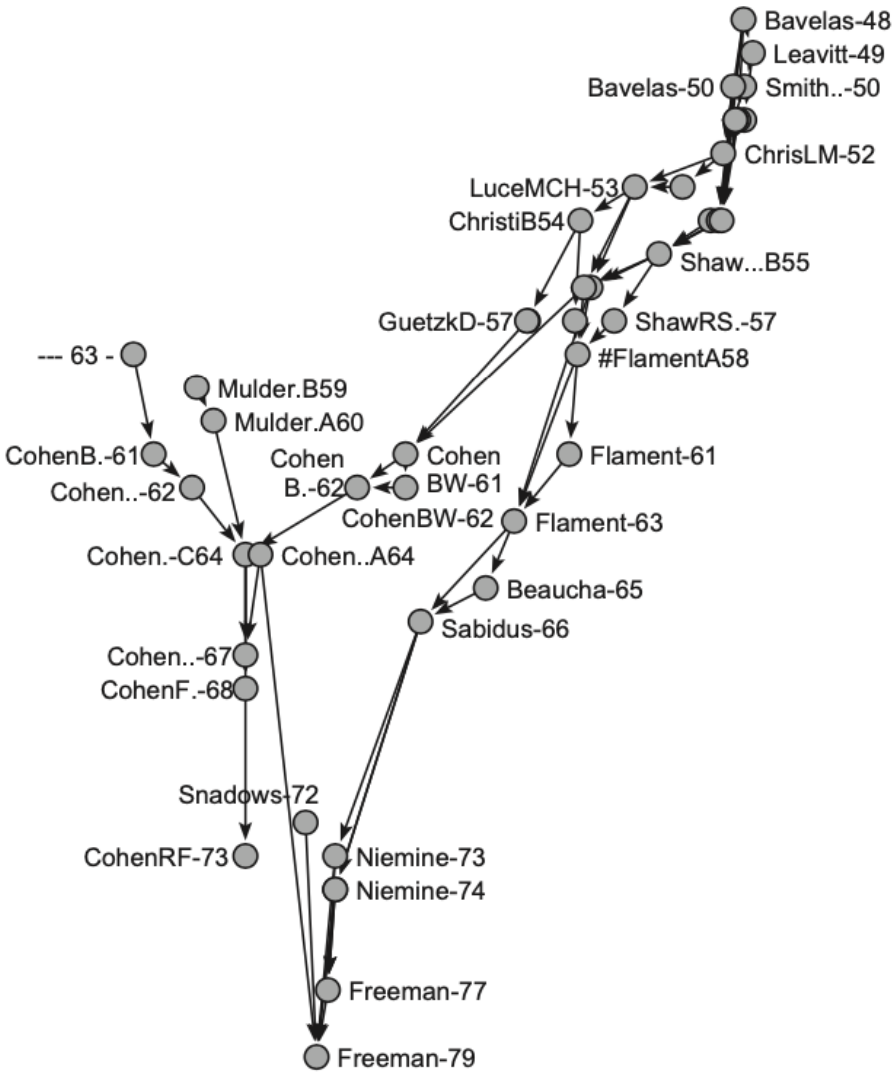


Figure 107. Main path component of the centrality literature network (not all names are shown here).

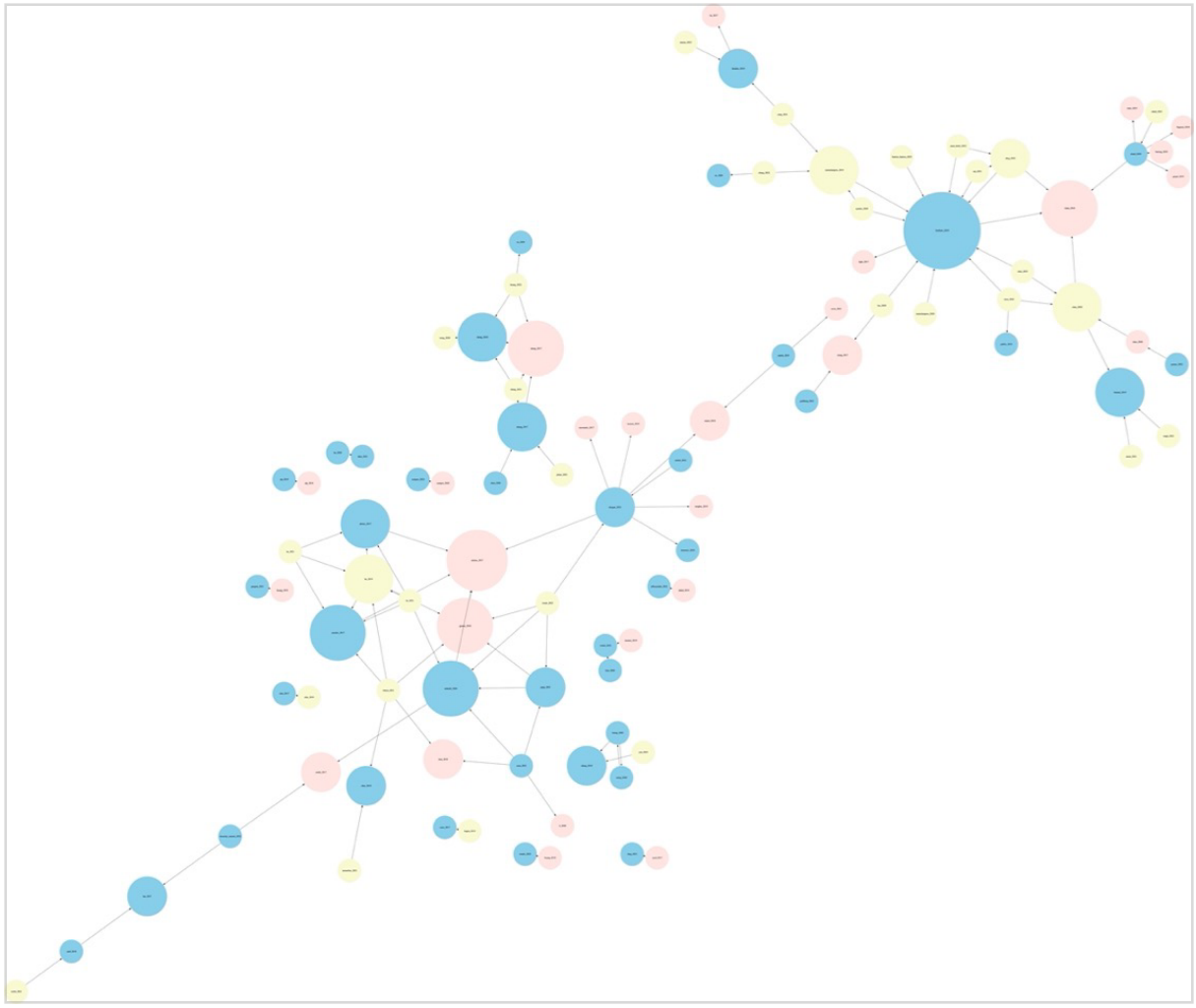
Limitation

The information stored in the .bib files revolves **solely around the initial set of selected papers**:

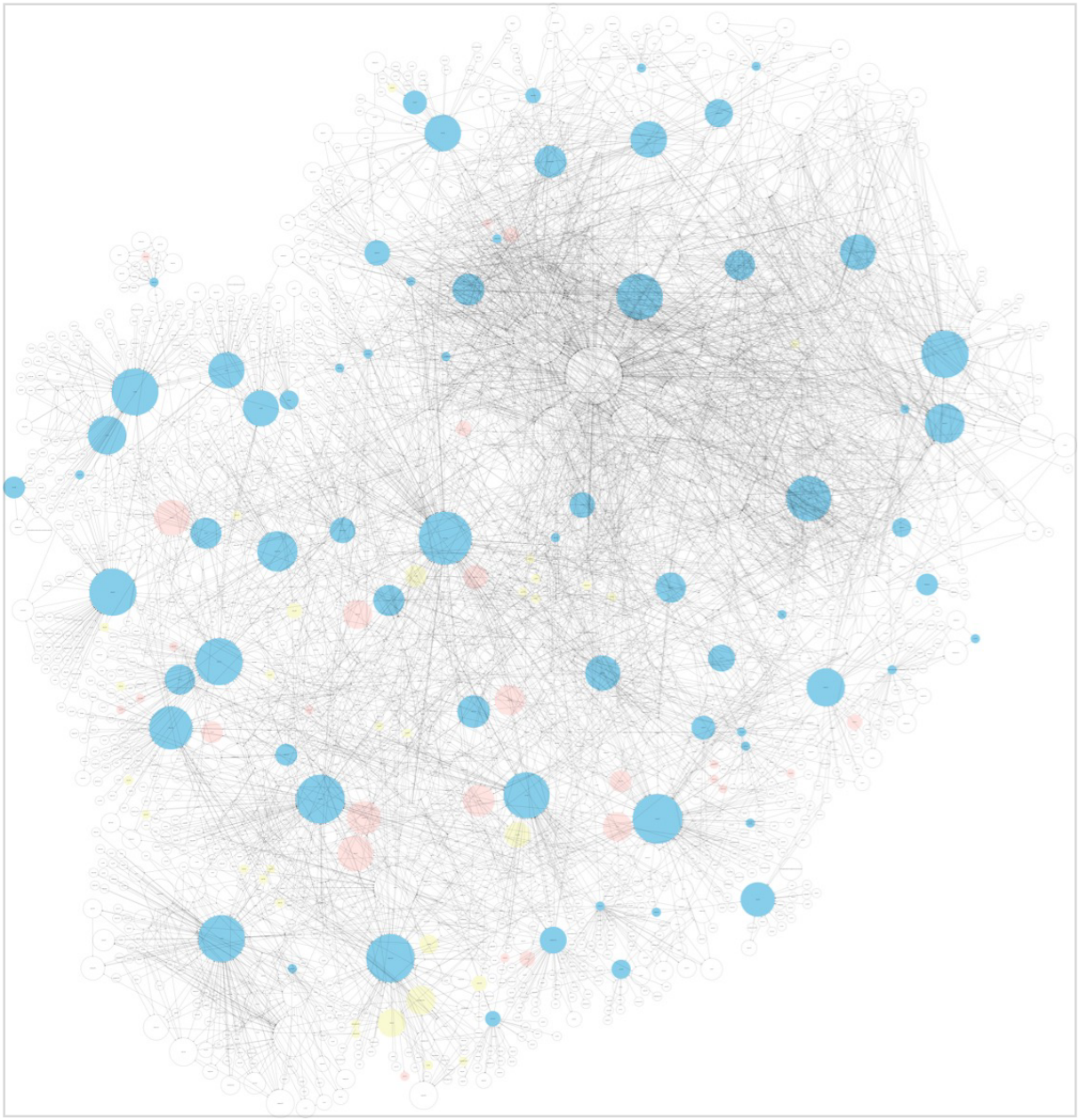
- Backward snowballing papers generate the outgoing edges
- Forward snowballing papers represent generate the incoming edges

Missing citations between the Forward / Backward Snowballing papers, which is required to compute more meaningful metrics.

1. Can we extend this such that BS and FS papers are **also connected with one another**?
   Yes, by extracting references from all the snowballing papers.
   Result



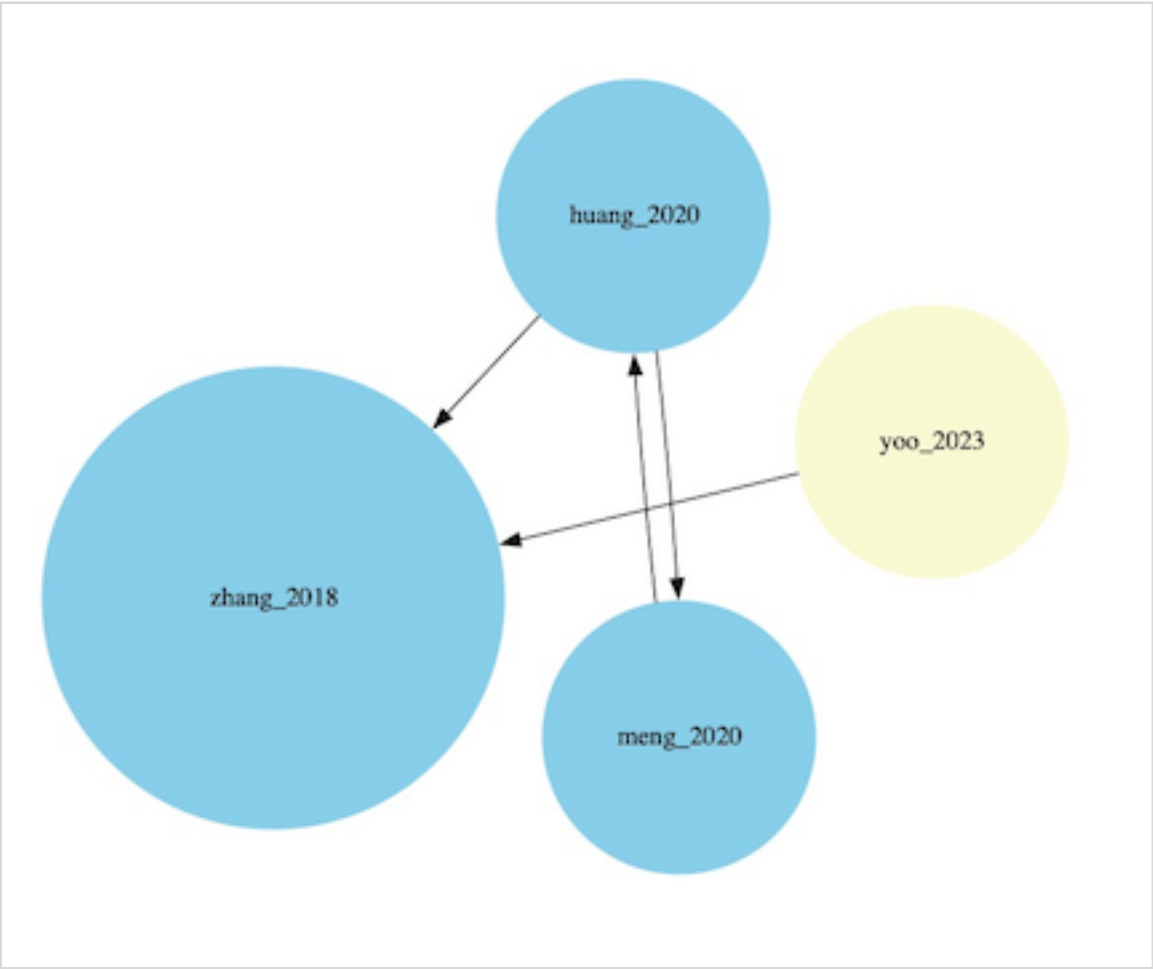
2. Can we do the same for the full graph containing **all** the analysed papers?



Containing the initial set of 65 papers and all retrievable FS and BS papers.

An interesting note (from Chapter 11)

"In principle, articles can cite only articles that appeared earlier, so the network is acyclic. Arcs never point back to older articles just as parents cannot be younger than their children. However, there are usually some exceptions in a citation network: articles that cite one another (e.g., articles appearing at about the same time
and written by one author)"



Data collection process and data items

Data items ➜ From Kitchenham (2004):

- The objective of this stage is to design **data extraction forms** to accurately record the information researchers obtain from the primary studies.
- The data extraction forms must be designed to collect **all the information needed to address the review questions** and the study quality criteria.
  - In most cases, data extraction will define a set of numerical values that should be extracted for each study.
  - Numerical data are important for any attempt to summarise the results of a set of primary studies and are a prerequisite for meta-analysis.

Data collection ➜ From Kitchenham (2020):

- "Specify the **method** used to collect data […], including how many reviewers […], whether they worked independently, […], details of automation tools used in the process."
- List, define and justify all outcomes for which data was sought, explaining their relationship to the research questions
- Describe any assumptions made about any **missing or unclear information**.

- […] define any classification systems used to **categorise the data items** and confirm how the data item relates to the research questions.

Data extraction form				
Item nr.	Item	Description	RQ	Mandatory
1	Year	Publication year	-	
2	Author(s)	Publication author(s)	-	
3	Title	Publication title	-	
4	Venue	Publication venue	-	
5	Topic labeling	Topic labeling approach(es)	RQ1	
6	Focus	Primary / Secondary focus on topic labeling	RQ1	
7	Type of contribution	Established / Novel approach for topic labeling	RQ1	
8	Underlying technique	Technique / Algorithm on which the topic labeling approach is based	RQ1	x
9	Topic labeling parameters	Parameter names and values used for topic labeling	RQ1	x
10	Approach details	Details of the employed approach	RQ1	
11	Motivation	What was the main motivator behind employing the topic labeling step?	RQ1	
12	Topic modeling	Underlying topic modeling approach(es)	RQ2	
13	Topic modeling parameters	Parameter names and values used for topic modeling	RQ2	
14	Label	Label description (e.g. single- multi-word) and nr of candidate labels per topic	RQ3	
15	Label selection	Selection approach(es) for label candidates	RQ3	x
16	Label quality evaluation	Quality metric(s) for label evaluation	RQ3	x
17	Assessors	Number and details of the assessors involved in the selection and evaluation	RQ3	x
18	Domain	Domain(s) of interest	RQ4	
19	Corpus	Origin, format, shape and content of the corpus	RQ4	
20	Document	Format of individual documents in the corpus	RQ4	
21	Pre-processing	Pre-processing steps performed on documents	RQ4	

- Item 10 is somewhat tied to items 6 and 7. In this context, 10 will be more extensive for those papers that propose novel labeling approaches and for paper that have labeling techniques as their primary focus.
- Item 18 can conceptually allow to **cluster** the collected papers by domain (useful for visualisation purposes).

#Thesis/Weekly notes#