DELIVERABLE D2.3

Grant Agreement:	**824350**
Project acronym:	**OSCAR**
Project title:	***O*pen *S*cien*C*e *A*eronautic & Air Transport *R*esearch**
Funding Scheme:	**Coordination and support action**
Start date of project:	**2019-01-01**
Duration:	**30 Months**

Date of latest version of Annex I against which the assessment will be made:
V.1.0.0 dated 2019-07-30

Qualitative and Quantitative Content Analysis of Five Representative Consortium Agreement Models

Due date of deliverable: 2019-07-30
Actual submission date: 2020-04-07
Deliverable version: Final V.1.0

Lead partner for this deliverable: IRB
Lead partner for the related work package: IRB

Name, title and organisation of the scientific representative of the project's coordinator:

Dipl.-Ing. Gerhard Pauly
Fraunhofer Institute for Manufacturing Technology and Advanced Materials IFAM
Fraunhofer-Gesellschaft zur Förderung der Angewandten Forschung e.V.

Acknowledgements: Many thanks to Tina Klages, Agnes Grützner, Beatrix Gittig, Renaud Bissling, Valérie Hachette and Gerhard Pauly, who have contributed to earlier versions of this report.

Project co-funded by the European Commission within Horizon 2020, the EU Framework Programme for Research and Innovation (2014-2020)		
Dissemination Level		
PU	Public	✓
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC.	

Report Approval Status

	Name	Organisation Short Name, Department, Function	Date	Signature	Comments
Author(s)	Maga, Martin	IRB	2019-07-30 2020-04-07		
Approval(s)	Gittig, Beatrix	IRB	2019-07-30		
	Agnes Grützner	IRB	2020-04-07		
	Dr. Klages, Tina	IRB	2019-07-30 2020-04-07		
Authorization(s)	Dr. Klages, Tina	IRB	2019-07-30 2020-04-07		

List of Distribution

Name	Organisation Short Name, Department	Date	Type of Distribution[1]	Distributed Report Parts[2]		
				Cover Page and Summary[2]	Main Report[2]	Annexes[2]
All researchers of the OSCAR consortium who access the OSCAR Content Server			D	✓	✓	✓

Explanation of notes and list of distribution:

[1] Type of Distribution: please use only the following codes

S = Originally signed print-out

P = Paper copy

D = Digital file

[2] Distributed Report Parts: please cross mark if applicable

Cover Page and Summary

Main Report = The whole report including cover page and summary with details, but no annexes or appendices

Annexes = All annexed separate documents

All data, statistics and media used or generated by this analysis can be found in the related zip files: "data.zip", "stats.zip" and "media.zip".

Abstract

Background: The OSCAR project aims (a) *to research the current state of open science in the European aeronautics and air transport (AAT)* research landscape and (b) to implement open science into the European AAT research landscape. To reach the second goal of OSCAR we strive (b.1) to develop an open science code of conduct for the European AAT research landscape and (b.2) to harmonise the main topics of this open science code of conduct with the common consortium agreement models (CAMs) in this field.

Objective: The primary objective of the present analysis is to establish an information basis for the strategic alignment of the project. The main question of the analysis at hand is as follows: *Is open science (already) relevant in the existing CAMs in European AAT research fields?*

Methods: We performed a qualitative and quantitative (multi variate) content analysis (Mayring 2014; Blasius and Baur 2014) of five representative CAMs used in the European AAT research landscape. Content analysis is a well-established, scientific method of empirical social science for objective information retrieval (Blasius and Baur 2014). The analysis is comprised of two main steps. *The first step* was to perform a theoretical background analysis in combination with an automatic topic modelling of open science to determine the important categories of open science. *The second step* was to analyse the content of the CAMs (a) qualitatively and (b) quantitatively.

Results: We determined 18 important categories of open science in general. The inter rater reliability between coder 1 and coder 2 is $\alpha = 0.422$ (Krippendorff's α). Although we did *not* achieve the level of agreement of $\alpha \geq .667$, there *is* however, a *systematic agreement between the two coders*, as shown by the Kendall rank correlation coefficient τ with $\tau = 0.344626$ ($p > .001$). The observed category frequencies by coder 1 are significant ($p < .001$) the ones observed by coder 2 are *not* significant ($p > .6$) (Fisher's exact test). The category frequency is significantly higher than expected ($p > .001$) (Fisher's exact test). The documents showed a high degree of inter-document similarity. Our analysis showed that the most relevant categories present in the CAMs are: (1) Intellectual Property, (2) *Open Source Software*, (3) *Open Data*, (4) *Ethics and responsibility*. However, it is interesting to note that *digitalisation* does not seem to be particularly relevant in the given CAMs.

Conclusion: Our analysis of the five major CAMs that are widely used in the European AAT research landscape shows that *open science and its underlying conceptual framework is indeed relevant in these CAMs*. In the forthcoming course of the OSCAR project, we should focus on developing communication strategies that tie on the four identified categories. The OSCAR project should focus on already used EU standards and guidelines and existing best practices of the industry and scientific community. To raise awareness, we should address the policy makers as well as the main stakeholders. Policy makers should make clear statements, commitments and rules. The main stakeholders should be well informed. We should develop simple opt-in, opt-out models for open science that can be used in the projects with ease. Our opt-in, opt-out models should emphasis the integration of conventional intellectual property management and open science practices.

Data: All data, statistics and media used and generated by this project can be found in the zip files: data.zip, stats.zip and media.zip

Contents

1 Motivation and relevance of this deliverable for the objectives of OSCAR

The main goal of the OSCAR project is to foster and implement open science in the European aeronautics and air transport (AAT) research landscape. This main goal of the OSCAR project has the following three related key components:

1. Analyse the status quo of open science in the European AAT research landscape.
2. Develop an open science code of conduct (CC).
3. Integrate or harmonise the open science conduct with the established consortium agreement models (CAMs).

The present analysis (deliverable D2.3) is one of the first steps to achieve the aforementioned grand goal (with its three key components) of the OSCAR project. In particular, the analysis at hand addresses the first key component, i.e., analysing the status quo of open science in the European AAT research landscape. One of the key elements of the AAT landscape in turn are the established CAMs. The analysis at hand deals with the *content of these CAMs*. In particular, the analysis at hand has the following two related purposes:

1. To provide information on the status quo of open science within the established consortium agreement models (CAMs) and
2. To strategically aligning the OSCAR project on the basis of the results of 1.

Based on the results of the analysis at hand the OSCAR project can perform informed steps towards its aforementioned main goal. In our analysis, we focus on the following two related questions:

1. Is open science already relevant in the CAMs?
2. Depending on the answer of 1., to which extend is open science a topic in the CAMs?

To answer these two questions, we performed a scientific *content analysis* of the five widely used CAMs in the European AAT research landscape. The method of *content analysis* (Mayring 2014) *is a de facto standard of empirical social science* (Bhattacherjee 2020). We combine two individual sub-methods of content analysis namely the *qualitative* and *quantitative* method of content analysis (Mayring 2014; Blasius and Baur 2014). This combined method of qualitative content analysis on the one hand and quantitative content analysis on the other hand is called *multivariate content analysis* (Blasius and Baur 2014). This approach, as opposed to pure reading—which is arguably not a scientific method in its own right—, enables us to conduct an *objective scientific analysis* of the CAMs *in the first place*. Please see the section: *Methods* for a more detailed elaboration on the chosen methods.

2 General theoretical background assumptions and agreement on used terminology

We are aware of the existence and importance of the philosophical and linguistic background of the current analysis, in particular of the philosophy of language, formal semantics and computational linguistics. Yet, the definitional and explanatory study of those underlying fundamental concepts and theories used in our analysis obviously goes beyond the scope of the present report. On the one hand, even a shallow investigation on those concepts and theories would be inappropriate to the facts. On the other hand, such a study would not be expedient regarding the main purpose of the report at hand. For these reasons, we elaborate only on the most necessary clarifications and background assumptions. In favour of the simplest possible argumentation, we will presuppose scientific common sense wherever possible and we will only give relevant elaboration where necessary.

2.1 Categories and concepts

The purpose of the analysis at hand is to obtain a clear, *objective* understanding of the content of the given CAMs. To ensure the greatest possible degree of objectivity, we treat the given CAMs as available *primary raw text data* in the course of our analysis. To analyse those text data we *use established methods of empirical social research*, in particular *content analysis* by Mayring (2014). One important element of content analysis (particularly of qualitative content analysis) are categories (Mayring 2014, 37). (We will elaborate on what qualitative and quantitative content analysis is in the *Methods* section.) Categories and

> "[c]oncepts are the building blocks of thoughts. Consequently, they are crucial to such psychological processes as categorization, inference, memory, learning, and decision-making. This much is relatively uncontroversial. But the nature of concepts—the kind of things concepts are—and the constraints that govern a theory of concepts have been the subject of much debate." (Margolis and Laurence 2019)

We cannot go into the language-philosophical background discussion regarding the nature of categories or concepts in the present analysis. Fortunately, for our analysis we only need the standard methods of linguistics, philosophy of language and empirical social science and only their essential notions namely words, hypernyms, hyponyms and synonyms (Jurafsky and Martin 2019). Please see the *Methods* section for more details on the methods used.

2.2 Frequencies, words, terms, types, token

For our *quantitative* content analysis, i.e., *frequency and probability analysis*, we use words as the most fundamental unit of analysis. In written language, words are groups of letters and can be considered the *smallest meaningful (semantic) unit*. Phrases are groups of words. For simplicity, we use the word "term" and the word "word" as synonyms. For counting purposes, it is useful to distinguish types from tokens (Jurafsky and Martin 2019). We call the individual instance or occurrence of a word a *word token* (Jurafsky and Martin 2019). We call the class of individual instances of a word a *word type*. (Jurafsky and Martin 2019) For example, in the sentence "To Be or Not to Be" there are 6 words, 4 types ("or", "be", "to" "not") and 2 tokens of the type "to", 2 tokens of the type "be", 1 token of the type "or" and 1 token of the type "not". For our frequency counts and estimates, we use both types and tokens depending on the respective purpose. Please see the *Methods section* for more details on the methods used.

2.3 Explicit and implicit content

Explicit text content is content that is written down directly or literally in a text. For example, the sentence "We commit ourselves to publish all our texts as open access publications." refers *literally* to *open access*. However, not all content is explicitly mentioned in all texts. For example, the sentence: “Transparency and reproducibility are virtues of science.” is *not* literally talking about *open science*. However, transparency and reproducibility are of course also important categories of open science. To capture the implicit, hidden or latent content of the given CAMs we use a well-established methodology stemming from natural language processing, namely *semantic network analysis* (Jurafsky and Martin 2019). In particular, use hypernyms (that is words that subsumes a number of other words in its meaning) and hyponyms (that is words whose meaning is enclosed by a hypernyms) and synonymy (words that have the same meaning) (Jurafsky and Martin 2019). For example, the word “mammal” is a hypernym and includes the meaning of the words “dog” and “cat”. For example, the phrase “helicopter” is a hyponym relative to the hypernym “aircraft” that captures its meaning as well as the meaning of “airplane”. We use this fact for our content analysis in the following way. The meaning of umbrella (collective) concepts is determined per definition by their subordinate concepts, i.e., by their hyponyms. *We consider the concept of open science to be a collective concept*. That means, one can refer to open science *explicitly* (directly) by using the phrase “open science” or *implicitly* (indirectly) by using hyponyms of open science like “open access” and “open source” etc. Please see the *Methods section* for more details on the methods used.

3 Methods

3.1 The methodological challenge

The methodological challenge we have to face in this analysis is the following: To answer the two aforementioned main questions, we need to determine the contents of the given CAMs. At this point, the question arises how we can gain knowledge about what the actual contents of the given CAMs are. The mere reading of the texts is arguably *not* an adequate scientific method. This is because the content supposedly made accessible by pure reading of a text is obviously almost *completely subjective* and *not* objective. Different readers have different opinions about what the content of a particular text is. Especially legal texts, like the contract templates (models) at hand, have a very specific status regarding their (legal) content. On the one hand, the authors of those contracts make the highest demands on accuracy with regard to the wording and content of these contracts. On the other hand, these contracts are *de facto* subject to constant interpretation—at the latest when a legal dispute arises. After all, an entire field of law, namely contract law, is devoted to the precise formulation, legal interpretation as well as handling of contract closing and breach of contract of contracts like the ones at hand. The discussion of the legal status of contracts, the legal issues or contract law in general is beyond the scope of the current analysis. *The crucial point is that contract models are in any case no exception to the rule regarding their subjective interpretability*.

3.2 Our approach

To solve the aforementioned methodological challenge of objectively gain knowledge about the contents of the CAMS at hand, we performed a scientific *content analysis* (Mayring 2014) of those CAMs. The method of content analysis (Mayring 2014) *is a de facto standard of empirical social science* (Bhattacherjee 2020). Content analysis is a method for intersubjective and systematic retrieval of content in data and texts (Mayring 2014; Blasius and Baur 2014). We *combine* two separate techniques of content analysis namely the *qualitative* and *quantitative* technique of content analysis (Mayring 2014; Blasius and Baur 2014). This combined techniques of *qualitative* content analysis on the one hand and *quantitative* content analysis on the other hand is called *multivariate content analysis* (Blasius and Baur 2014). This approach, as opposed to pure reading—which is arguably *not* a scientific method in its own right—, enables us to conduct an *objective scientific analysis* of the given CAMs *in the first place*. By utilising the well-established method of multivariate content analysis, it is possible to *objectively* assess the content in the available data, i.e., the available established CAMs

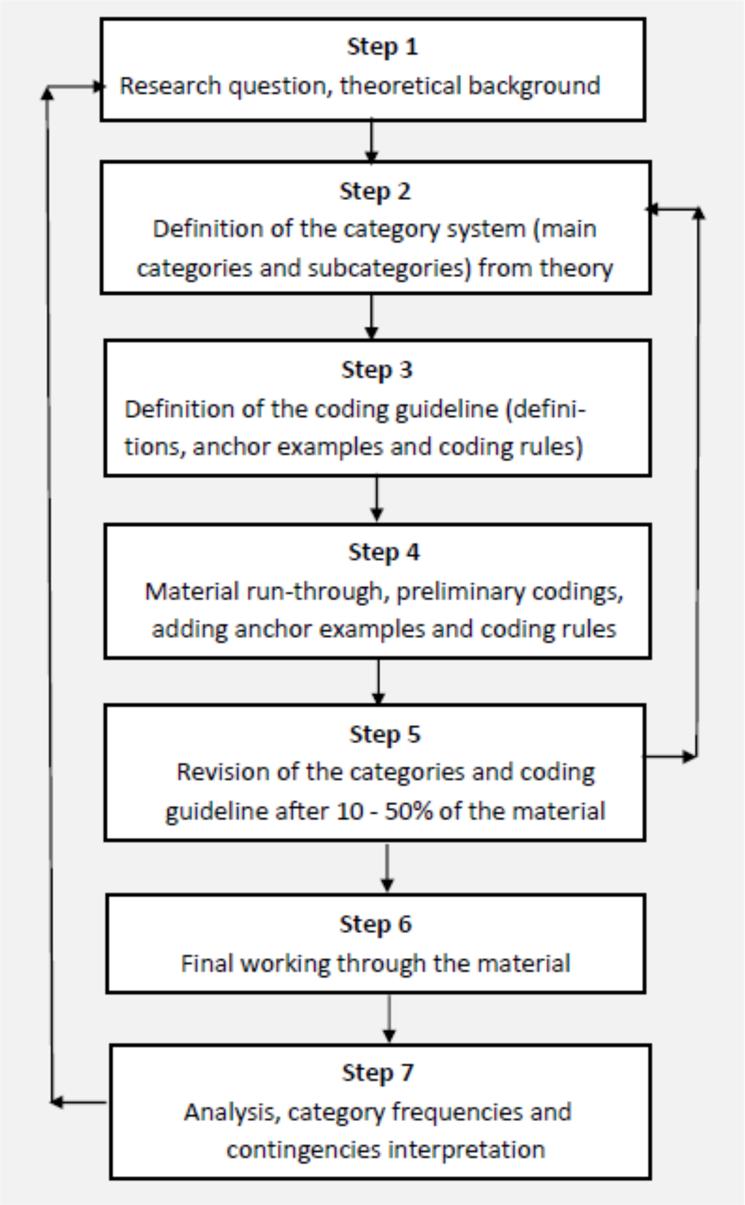
(Mayring 2014; Blasius and Baur 2014). We will elaborate on our approach in the next section: *Content analysis workflow*.

3.3 Content analysis workflow

3.3.1 General content analysis workflow

The general process of content analysis by Mayring (2014) is comprised of 11 main steps. In the initial steps, it is first defined which material (data) is to be analysed. Furthermore, the genesis of the data is described as well as its formal characteristics. In the next steps, the scientific question is differentiated, a corresponding analysis technique is developed and the units of analysis are determined. The core steps consist of developing a good system of categories, examining the material at the level of the selected analysis units and consolidating the results. In the final steps, the results are statistically evaluated and interpreted. Please see Mayring (2014) for more details on the general procedures and methods and Figure 1 for an overview.

Figure 1: Steps of deductive category assignment; Source: (Mayring 2014, 96)

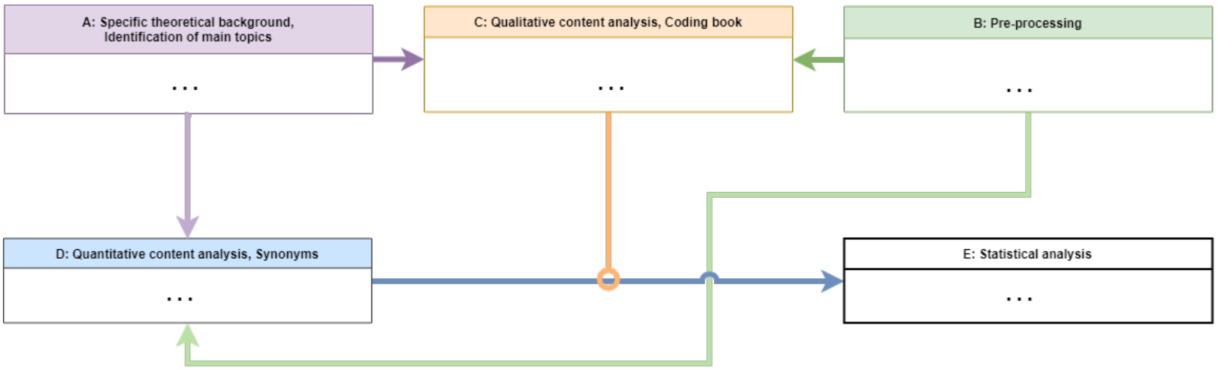

Step 1
Research question, theoretical background
Step 2
Definition of the category system (main categories and subcategories) from theory
Step 3
Definition of the coding guideline (definitions, anchor examples and coding rules)
Step 4
Material run-through, preliminary codings, adding anchor examples and coding rules
Step 5
Revision of the categories and coding guideline after 10 - 50% of the material
Step 6
Final working through the material
Step 7
Analysis, category frequencies and contingencies interpretation

There are different ways of implementing the *core steps* of qualitative content analysis, i.e., the *development of the category system*. We choose the method of *deductive category assignment* (Mayring 2014, 95–103). In general, this method consists of seven steps. The first step is to elaborate the theoretical background of the question or the topic. This theoretical background can then be used to develop and define the according categories in the second step. The third step is to create a *coding book* consisting of the following parts: The coding book contains all *categories* that have been worked out before. For each category, a *coding rule* is defined, which specifies which passages of the material fall under this category. For each category, there is an *anchor example* from the material, including a reference to where it was found. In the fourth step, one or more persons, called coder, systematically runs through the material and records passages that fall into one of the previously defined categories according to the rules of the coding book. In this step, after 10% to 50% of the material has been processed, the coding book can be revised if systematic errors occur. (We have omitted this step in our analysis.) The seventh and last step is to statistically evaluate the frequency of the categories and interpret the results. Figure 1 gives an overview of this method. For more details on the deductive method of qualitative content analysis, please see Mayring (2014).

3.3.2 Our specific content analysis workflow

Our specific content analysis workflow is closely based on the approach by Mayring (2014), described above. Our content analysis workflow is comprised of *five main steps, A, B, C, D and E*. For an overview of these steps, please see Figure 2.

Figure 2: Overview of our analysis workflow


A: Specific theoretical background, Identification of main topics
…
C: Qualitative content analysis, Coding book
…
B: Pre-processing
…
D: Quantitative content analysis, Synonyms
…
E: Statistical analysis
…

3.3.2.1 Step A

The first *step A* consists of a specific *theoretical background analysis and the identification of main topics of open science*. The results of step A, i.e., the 18 important categories of open science provide the starting point for the next step B.

3.3.2.2 Step B

The second *step B* consists of the *pre-processing of the given CAMs*. This has the purpose of anonymising the documents and making the documents easily readable in the PDF for the coders. We were handed eight documents. These eight documents represented different variants of only five CAMs in total. Some of the documents contained additional information about the same CAM, such as comments or extra sections. This means that there were redundancies, which we removed carefully. The results of step B, i.e., six clean CAM PDF documents provide the starting point for the next steps C and D.

3.3.2.3 Step C

The third *step C* consists of the *qualitative content analysis including the creation of the coding book*. The results of step C, i.e., CAM corpus category frequency table A (in Figure 3 and 4) together with the results of D provide the starting point for the last step E.

3.3.2.4 Step D

The fourth *step D* consists of the *quantitative content analysis including the creation of the category model via synonyms*. The results of step D, i.e., CAM corpus category frequency table B (in Figure 3 and 4) together with the results of step C provide the starting point for the last step E.

3.3.2.5 Step E

The fifths and last step E consists of the *statistical analysis of the qualitative and quantitative content analysis, step C and D*.

Each of those five main steps in our analysis workflow is comprised of several sub-steps. For a detailed view of our analysis workflow, please see Figure 3.a and Figure 3.b as well as the complete, scalable versions the workflow representation in the annex. After we have roughly described our analysis workflow, we describe the individual steps in detail in the following sections.

3.3.3 Specific theoretical background of open science

This step (step A) deals with the specific theoretical background analysis and identification of important categories of open science. This step consists of two independent parts:

1. Qualitative analysis of standard texts on open science and
2. Quantitative text mining of standard texts on open science

These two parts serve the purpose of identifying the main categories of open science *as objectively as possible*. In the *literature analysis part*, we systematically research and read a selection of standard texts on open science and excerpt standard definitions. From these excerpts, we create a list of the main principles, topics, concepts or categories of open science. In the *text mining part* we algorithmically scan a standard text corpus and algorithmically extract a list of categories via *automatic topic modelling* (Wikipedia 2020c). Both independently generated lists of six main categories of open science each are merged together where possible or recombined via abstraction to new categories. The result is a single list of 18 categories. This list of is used in the creation of the coding book. Please see Figure 3, A.1 and A.2.

3.3.4 Pre-processing the given CAM documents

This step (step B) deals with the preparation, sorting and the pre-processing of the given CAM documents. The purpose of this step is to systematically run through the primary raw text data (CAMs) and decide whether a given document will be used or whether there are bad redundancies. The result is a clean set of *secondary text data* (CAMs text corpus) that can be used for the next steps in our analysis. Please see Figure 3, B.

3.3.5 Qualitative content analysis

This step (step C) in our content analysis workflow deals with the qualitative content analysis including the creation of the coding book. The goal of this step is to arrive at an *inter-subjective category frequency table*. This step consists of two independent parts:

1. Creation of the coding book and
2. Run-through with two independent coders (persons).

In the first part, we use the list of categories identified in our previous specific theoretical background analysis to create the coding book (see Table 12 in the annex). The coding book is a table and is comprised of four columns (parts): The first columns consist of all categories that we have obtained from the previously created list of important categories. The second column contains a definition for each category. The category definitions are determined by their canonical dictionary definitions (Lexico.com and Oxford University Press (OUP) 2019). The third column contains rules for each category. Each rule specifies the conditions under which a text passage falls under the according category. The fourth column contains anchor examples for each category found in the text corpus. Please see Figure 3, B1.

In the second part, we run through the CAMs text corpus with two independent coders and the help of the previously created coding book. A coder is a person who systematically goes through the text corpus with the help of the coding book and copies all the found references according to the rules of the coding book and enters them with a corresponding reference into a table of his own. Please see Figure 3, B.2. We will discuss this method in more detail later in the main section: *Qualitative content analysis*.

3.3.6 Quantitative content analysis

This step (step D) deals with the quantitative content analysis including the creation of a category model via synonyms. The goal of this step is to arrive at an *objective category frequency table*. This step is non-trivial in nature because it is very hard to determine the *priori* base frequencies of category occurrence. We tackle this challenge by a rather simplistic but effective category model. We build our model by utilising the assumption of hypernyms and synonyms, which is also a silent background assumption in qualitative content analysis (see section: General theoretical background). For each category, we determine a list of synonyms with the help of a standard dictionary (Lexico.com and Oxford University Press (OUP) 2019). We then consult a comprehensive word frequency list (Word frequency data 2019) to determine the base frequency of each synonym. From these frequencies, we can calculate the probabilities and expected values for each synonym. We model each category probability by the combined probabilities of the corresponding synonyms. Finally, we search for all synonyms for each category in the CAMs text corpus and count their occurrence frequencies. This procedure allows us to specify the a priori expected frequency for each category against which we can test the observed category frequencies of each coder. Please see Figure 4, D. We will discuss this method in more detail later in the main section: *3.5 Quantitative content analysis*.

3.3.7 Statistical analysis

This step (step E) deals with the statistical analysis of the results from the qualitative and quantitative content analysis, i.e., the CAM corpus category frequency table A and B (see Figure 3 and 4). We use the classic Fisher's exact test (Wikipedia 2019) to calculate the significance level for each coder and the category frequencies. We use the Krippendorff's alpha (Krippendorff 2011) and the Kendall rank correlation coefficient (Abdi 2007; Wikipedia 2020b) to calculate the inter coder reliability. The statistical results can then be interpreted and this interpretation can be used as a basis for drawing conclusions with respect to the project. Please see Figure 4, D.

Figure 3: Detailed analysis workflow part 1

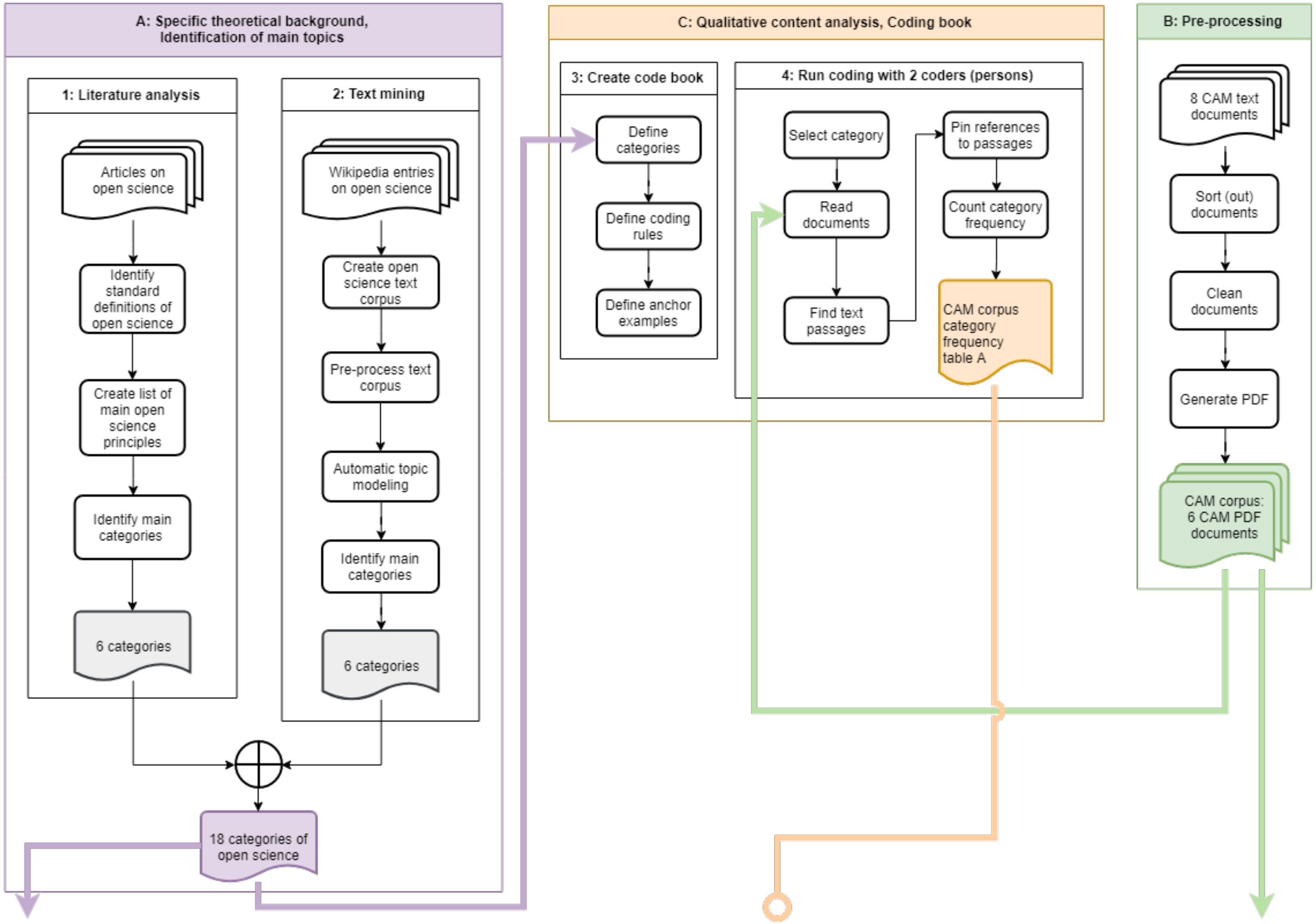
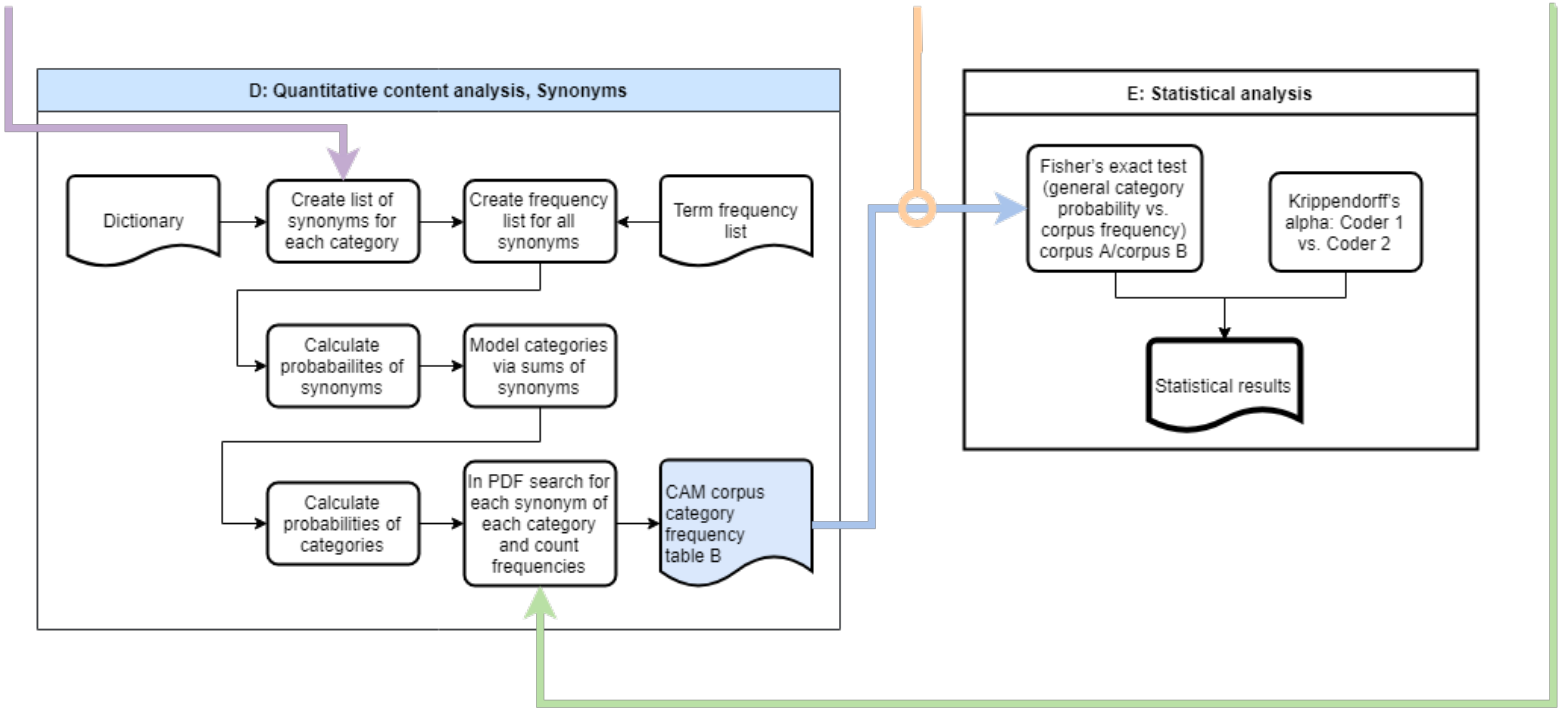

A: Specific theoretical background, Identification of main topics
1: Literature analysis
Articles on open science
Identify standard definitions of open science
Create list of main open science principles
Identify main categories
6 categories
2: Text mining
Wikipedia entries on open science
Create open science text corpus
Pre-process text corpus
Automatic topic modeling
Identify main categories
6 categories
18 categories of open science
C: Qualitative content analysis, Coding book
3: Create code book
Define categories
Define coding rules
Define anchor examples
4: Run coding with 2 coders (persons)
Select category
Read documents
Find text passages
Pin references to passages
Count category frequency
CAM corpus category frequency table A
B: Pre-processing
8 CAM text documents
Sort (out) documents
Clean documents
Generate PDF
CAM corpus: 6 CAM PDF documents

Figure 4: Detailed analysis workflow 2


D: Quantitative content analysis, Synonyms
Dictionary
Create list of synonyms for each category
Create frequency list for all synonyms
Term frequency list
Calculate probabailites of synonyms
Model categories via sums of synonyms
Calculate probabilities of categories
In PDF search for each synonym of each category and count frequencies
CAM corpus category frequency table B
E: Statistical analysis
Fisher's exact test (general category probability vs. corpus frequency) corpus A/corpus B
Krippendorff's alpha: Coder 1 vs. Coder 2
Statistical results

4 Research question and differentiation of the hypothesis

The central question of the current deliverable arises directly from the main aim of the OSCAR project itself. *Is it possible to integrate or harmonise the statements of commitment to some of the major open science principles into some CAMs commonly used in the European AAT research landscape?* Arguably, it is possible to integrate open science into the European AAT research landscape in general—at least we are not aware of any compelling argument that would prove that it is impossible. The *bona fide* possibility leads directly to the follow-up question: *How* exactly can open science being integrated into the AAT CAMs? To answer this question, it is necessary to analyse the contents of representative CAMs with regard to open science. The instrumental normative objective conditional in the context of this question is as follows:

> **I**: If open science is *implicitly relevant* in some CAMs, then a strategy for integrating open science into these CAMs should be pursued that exploits the respective fact of *existing or not existing conceptual frameworks for open science*.

The existence or non-existence of a conceptual framework for open science places different demands on our upcoming strategic steps of the OSCAR project. Therefore, it is important to answer this question in advance of any further action or decision. To see whether the above implication is *true*, we need to define the antecedence *viz* what it is for open science to implicitly relevant in the CAMs in the first place. Only then, we can decide whether the antecedence of **I** is true. Only then, in turn, we can take appropriate, i.e., informed actions regarding the overarching goal of the OSCAR project. The following conditional working definition of relevance (antecedence of **I**) of open science in the given CAMs at hand is sufficient for the purposes of the current analysis:

> **D**: If more than half (50%) of the important categories of open science occur significantly frequent in the CAMs, then open science is implicitly relevant in the CAMs.

The assumption of this conditional definition is twofold:

First, for the purpose of this paper, we make the fundamental assumption that open science is a collective term (see section *2 General theoretical background assumptions and agreement on used terminology*) that gathers a non-exhaustive list of activities or paradigms, i.e., categories of open science that are loosely associated with each other. Examples of such categories, activities and paradigms are open access, open data, open source, etc. In this sense, we not forced to give necessary or sufficient conditions of the concept of open science, but only a list of loosely connected categories of open term (see section *2 General theoretical background assumptions and agreement on used terminology*). Those categories (sub-categories of open science) can then be systematically measured in the given texts (see section *3 Methods*).

Second, we assume that the occurrence frequency of a category in a text, which *cannot* be explained by pure chance, is a sufficient condition of *explicit* relevance of that category itself. Furthermore, analogously we assume that the *collective* occurrence frequency of the respective category synonyms or hyponyms in a text, which *cannot* be explained by pure chance, is a sufficient condition for *implicit* relevance of the respective collective category. Please refer to section *2 General theoretical background assumptions and agreement on used terminology* and section *3 Methods* for further information.

The core idea of the working definition **D** then is that a collective category occurs implicitly in a text when more than 50% of its synonyms or hyponyms, i.e., adjacent categories or sub-categories occur in that text. If all pre-determined sub-categories of a collective category occur in a text, then the respective category is almost surly implicitly relevant in that text. Vice versa if none of the sub-categories of a respective collective category is present in a given text, then the collective category almost surly is *not* implicitly relevant in that text. The threshold for a collective category to be implicitly relevant is *a priori* 50% of the occurrences of the corresponding sub-categories. To determine that more than 50% of sub-categories occur in a given text, a provisionally fixed list of sub-categories is required. This in turn, requires us to determine a fixed set of important sub-categories open science.

For example, if want to know if the category “aircraft” is implicitly relevant in given text we first determine a list of important sub-categories like: “airplane”, “helicopter”, “zeppelin”. We fix this list as a first approximation to our category, i.e., we assume that this list contains all the important sub-categories of the collective category “aircraft”. If we find that only “zeppelin” appears in our text, we cannot claim that, the category aircraft is implicitly relevant. If we find however, that both “zeppelin” and “helicopter” appear in our text, we can be almost sure that “aircraft” is indeed an implicitly relevant category in our text.

Having the working definition **D** and a fixed list of important categories (see section *5.1 Specific theoretical background of open science, identification of main categories of open science*) we are able to formulate our *hypothesis* (alternative hypothesis) and our *null hypothesis* as follows:

> **H0**: Open science is *not* implicitly relevant in the Consortium Agreement Models.

> **H1**: Open science *is* implicitly relevant in the CAMs.

We test our hypothesis pair by utilising a combination of a qualitative content analysis and quantitative analysis in the way described in detail the following sections. For more details, see also section *3 Methods*.

However, before we can run our qualitative and quantitative content analysis, we first need a *fixed list of sub-categories of open science*. To arrive at a reasonable fixed list of categories of open science we focus on determining only the most important categories of open science. In the next section, we describe the process of determining which categories of open science are important. Based on our experience with open science we first performed a theoretical analysis of standard texts on open science. Second, we performed a quantitative automated topic modelling of a representative text corpus. By combining the results of both analysis methods, we could derive a reasonably good approximation of a fixed list of important categories of open science.

5 Content Analysis

5.1 Specific theoretical background of open science, identification of main categories of open science

5.1.1 Qualitative analysis of standard texts of open science

Open science can be understood in a narrow sense and in a broader sense. In a narrow sense, open science denotes a *certain way* of doing science. In this narrow sense, open science can be understood as a specific scientific activity. In a broader sense, open science denotes a scientific-political movement (Fecher and Friesike 2014). In this broader sense, open science can be understood not only as a scientific activity but also as schools of thought that have a certain agenda (Fecher and Friesike 2014). In both cases, open science focuses on the sustainable opening of as many dimensions and aspects of science as possible to as many people as possible (Bezjak et al. 2018). Important fundamental principles of open science are freedom, transparency, reproducibility, reusability and open communication (Bezjak et al. 2018).

Open science as an activity: Different authors define open science differently. In general, open science can be understood as a certain type or kind of science that follows certain criteria or principles that can be formulated as rough and ready rules. Important principles among others are (Bezjak et al. 2018; Wikipedia 2020a):

1. **Open Access** Make your scientific publications freely available. There are different strategies for publishing according to the Open Access principle, two of them are: Gold Open Access and Green Open Access.
2. **Open Data** Make your research data freely available. (In the narrower sense, the concrete data records on which the published works are based on). See, for example, the FAIR principles (Findable, Accessible, Interoperable, Reusable).
3. **Open Source** Make the software that is created and used in the research process freely available.
4. **Open Notebooks** Make your scientific notebooks freely available.
5. **Open Peer Review** Make the peer review process openly available.
6. **Open Educational Resources** Make your teaching and training materials freely available.
7. **Open Methodology** Make your scientific methods freely available.
8. **Citizen Science** Open your research to people who are not full-time or professional scientists.
9. **Open Infrastructure** Make the infrastructure of your research freely available.
10. **Open Metrics** Make the metrics with which you measure the scientific impact freely available.

Open science as political movement: According to Fecher and Friesike (2014) open science can be understand as political movement. This political movement can be divided into five schools of thought (Fecher and Friesike 2014):

1. **The democratic** aim of this school of thought is to make knowledge accessible to as many people as possible.
2. **The pragmatic** goal of this school of thought is to open up the process of knowledge production.

3. **The infrastructural** goal of this school of thought is to develop platforms, tools and services that are freely available.
4. **Public Aim** of this school of thought is to make science freely available to citizens.
5. **Measurement** The aim of this school of thought is to develop freely available metrics.

Combining categories of both the narrow and the broad stance of open science described above, we can directly derive the following six preliminary categories:

Preliminary categories

1. **Open access** (availability, publications, …)
2. **Open data** (fair use and re-use, repositories, …)
3. **Open source** (software, source code, …)
4. **Open infrastructure** (platforms, processes, technology, guidance, …)
5. **Open knowledge and open knowledge transfer** (innovation, mutual exchange, collaboration, …)
6. **Open to public contribution** (citizen science, public aim, participation, …)

5.1.2 Quantitative automated topic modelling neutral text corpus on open science

So far, we have only elaborated which categories of open science are relevant in some standard texts on open science. The categories used are intuitively correct, yet they are inherently *qualitative*. In order to achieve a greater degree of objectivity regarding the categories of open science, we consider it important to make an additional quantitative categorisation. In this way, we can crosscheck qualitative and quantitative categories.

To crosscheck the categories we identified so far, we performed a quantitative topic analysis by querying the English Wikipedia with the search phrase “open science”. We have chosen Wikipedia to get the most neutral, i.e., non-biased and at the same time most informative corpus on the topic of open science.

For our quantitative automated topic analysis of the fetched corpus, we used the *Orange Data Mining and Machine Learning toolbox* (Orange 2019). Our automated topic modelling process consists of the following four steps.

5.1.2.1 First step: fetch Wikipedia entries

In the first step, we used the MediaWiki RESTful web service API to fetch 25 entries on the English Wikipedia that contain the phrase “open science” in their title, summary or content.

Text corpus

1. Document count:	25
2. Total types:	4.856
3. Total tokens:	27.709

5.1.2.2 Second step: Pre-processing

In the second step, we pre-processed the fetched raw text corpus in the following way:

1. Transforming: We transformed the entire raw text to lowercase, removed accents, and URLs.
2. Tokenization: We tokenized the raw text into single terms using the following Regular Expression: We filtered the raw text by applying the regular expression: \w+\g meaning match all alphanumeric word characters one or more times for the entire document.
3. Normalisation: We used the UDPipe Lemmatizer (Institute of Formal and Applied Linguistics 2019) to determine the stem, base or root form of the words.
4. Filtering: We filtered out noisy character sequences like numbers or stop words using the following Regular Expression: \.|,|:|;|!|\?|\(|\)|\||\+|'|"|‘|’|“|”|'|\’|…|\-|–|—|\$|&|*|>|<|\/|\[|\]|[0-9]\g.

The core idea of this step is to clean up the raw text data as much as possible without removing important information.

5.1.2.3 Third step: bag of words

In the third step, we generated a bag of words (an unordered list of all words disregarding grammar etc. but keeping multiplicity).

We used a sublinear term weighting to get the term frequency count (ML Wiki 2019). In particular we used term frequency–inverse document frequency (TF-IDF) with smoothing (add one) to reflect how relevant (or important) a word is to a document in our corpus (ML Wiki 2019).

Term frequency (TF) is just the raw count of the occurrences of a term (Jurafsky and Martin 2019, 105). Document frequency (DF) of a term is the number of documents a term occurs in. (Jurafsky and Martin 2019, 106). Inverse document frequency (IDF) of a term is equal to the number of documents divided by the DF of that term. (Jurafsky and Martin 2019, 107).

The core idea of this step is that a specific category occurs more frequently in a text than it would be by pure chance, if it were relevant to the author(s). If, for example, the word “open” occurs relatively frequent, then the category “openness” is a relevant category in the corresponding text. However, common terms like “the” are very frequent and are not important in the narrow thematic sense. That means that term frequency per se is an incorrect measure of relevance. In contrast, terms that are less frequent carry more information. We address this issue by applying a term frequency–inverse document frequency (TF-IDF) in this step. Please see Jurafsky (2019) for further information on these topics.

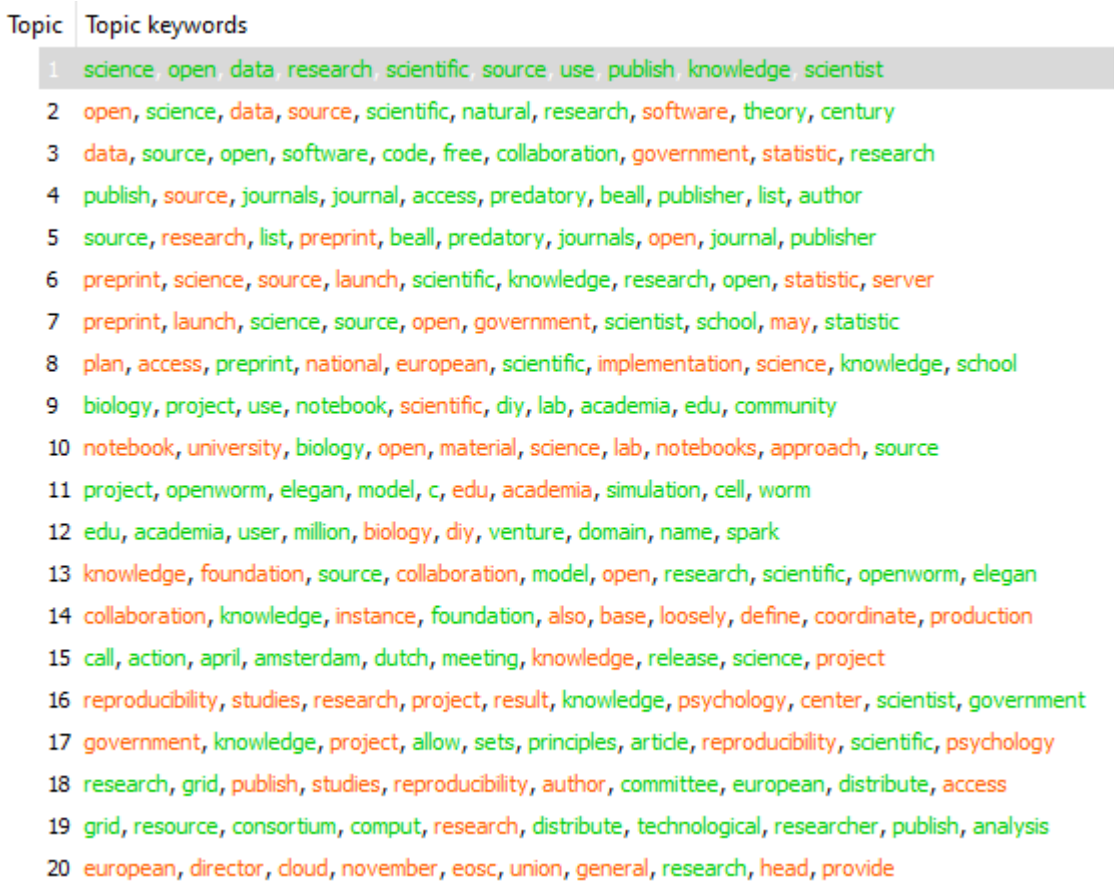
5.1.2.4 Fourth step: automated topic modelling

In the fourth step, we performed a statistical topic modelling using a latent semantic analysis (Wikipedia 2020b). In this method, the text corpus is represented by a matrix constructed via singular value decomposition containing rows of unique words and columns representing paragraphs. We measure the term similarity between column vectors with the most popular measure, i.e., the cosine similarity metric (Jurafsky and Martin 2019, 7). Figure 5 shows the top 20 topics generated by this model. Each topic is represented by a *nameless* row. Each row has a row number. This row number can be used to refer to a specific topic. Each row consists of adjacent keywords that give rise (computationally) to the topic. The rows are sorted by importance. The most common keywords corresponding to the most important topic are listed in the first row and this row is labelled with the number 5. A green keyword indicates that a text or text fragment that contains this keyword belongs to the corresponding topic. A red keyword indicates that a text or text fragment that contains this keyword does not belong to the corresponding topic.

The topics are represented by nameless rows. We choose the topic names by using the green keywords of the corresponding row. We selected the first six most important topics. As shown in Figure 5, the following six important topics can be identified:

1. **Open Science** in general understand as a collective term that includes open data, open access, open knowledge, et cetera.
2. **Science in general** as collective term that includes research and natural science but excludes open data and open source software.
3. **Open Source** including open source software, open code and collaboration but excluding open data and government.
4. **Open Access** including journals, publishers and authors but excluding open source.
5. Publication landscape including journals, publishers but excluding preprint and open access.
6. **Knowledge transfer** including open research, open knowledge excluding preprint, (open science) and open source.

Figure 5: Automated topic modelling



Topic	Topic keywords
1	science, open, data, research, scientific, source, use, publish, knowledge, scientist
2	open, science, data, source, scientific, natural, research, software, theory, century
3	data, source, open, software, code, free, collaboration, government, statistic, research
4	publish, source, journals, journal, access, predatory, beall, publisher, list, author
5	source, research, list, preprint, beall, predatory, journals, open, journal, publisher
6	preprint, science, source, launch, scientific, knowledge, research, open, statistic, server
7	preprint, launch, science, source, open, government, scientist, school, may, statistic
8	plan, access, preprint, national, european, scientific, implementation, science, knowledge, school
9	biology, project, use, notebook, scientific, diy, lab, academia, edu, community
10	notebook, university, biology, open, material, science, lab, notebooks, approach, source
11	project, openworm, elegan, model, c, edu, academia, simulation, cell, worm
12	edu, academia, user, million, biology, diy, venture, domain, name, spark
13	knowledge, foundation, source, collaboration, model, open, research, scientific, openworm, elegan
14	collaboration, knowledge, instance, foundation, also, base, loosely, define, coordinate, production
15	call, action, april, amsterdam, dutch, meeting, knowledge, release, science, project
16	reproducibility, studies, research, project, result, knowledge, psychology, center, scientist, government
17	government, knowledge, project, allow, sets, principles, article, reproducibility, scientific, psychology
18	research, grid, publish, studies, reproducibility, author, committee, european, distribute, access
19	grid, resource, consortium, comput, research, distribute, technological, researcher, publish, analysis
20	european, director, cloud, november, eosc, union, general, research, head, provide

5.1.3 Combining qualitative and quantitative results

By combining and rearranging the results from the quantitative topic modelling with the results from the qualitative theoretical analysis, we arrived at the following reasonable list of categories of open science. Both independently generated lists of six main categories of open science each are merged together where possible or recombined via abstraction to new categories. The result is a single list of 18 categories. Whenever possible we merged categories and/or choose a more abstract or general category to subsume similar categories under a new category. We choose the following 18 categories of open science to be the most important (see also Annex, Table 12):

1. Openness
2. Sharing / Giving / Contribution / Collaboration
3. Accessibility / Availability
4. Closed / Non-Disclosure / Confidentiality / Privacy / Restrictions / Limits
5. Public / Society / Community
6. Publication / Dissemination / Distribution / Deployment
7. Patent / Intellectual Property
8. Knowledge / Knowledge transfer
9. Value / Added value
10. Data

11. Repository
12. Governance
13. Quality / Interoperability / Standards / Practices / Best practices / Sustainability / Re-use / Transparency / Verifiability / Falsifiability / Visibility
14. Ethics / Fairness / Equality / Responsibility
15. Infrastructure / Platform
16. Copyright / Licensing
17. Digitalisation
18. Software / Source Code

It might be possible to collate the categories in a different way or to reduce it even more by abstraction but for the purpose of this report we consider this rather pluralistic list a sufficient approximation. Furthermore, the way in which we group the categories will have no effect on the results of our analysis.

5.2 Qualitative content analysis

5.2.1 Background, context and genesis of the objects under investigation

To support the European research area the European Commission created the *Framework Programmes for Research and Technological Development* (FP1-FP8). The last FP is called *Horizon 2020* and is funded with 77 billion Euro in the period of 2014-2020 (EU-Büro des BMBF 2019; Kooperationsstelle EU der Wissenschaftsorganisationen 2019; European IPR Helpdesk 2019).

In Horizon 2020, it is binding for all funded projects that the project partners of a project, i.e., the consortium to conclude a *Consortium Agreement (CA)* (EU-Büro des BMBF 2019; Kooperationsstelle EU der Wissenschaftsorganisationen 2019; European IPR Helpdesk 2019).

The CA is fixated in a contract under private law, concluded between the partners of a consortium in a research project. The CA regulates the internal relationship between the individual partners, i.e., the rights and obligations among themselves (EU-Büro des BMBF 2019; Kooperationsstelle EU der Wissenschaftsorganisationen 2019; European IPR Helpdesk 2019).

Despite the fact, that the European Commission enjoin the conclusion of a CA it does not however review the contents of the CA. Yet, the content of the CA is defined in the Roles of Participation. The European Commission gives only general guidance for developing a CA (EU-Büro des BMBF 2019; Kooperationsstelle EU der Wissenschaftsorganisationen 2019; European IPR Helpdesk 2019).

The conclusion of a CA is not mandatory or explicitly required by the European Commission. Yet, the CA between the coordinator and the beneficiaries is meant to define their rights and obligations and their governance as well concerning the implementation of the action. The CA contains rules of the grant agreement that contains and describes the *rules for participation*. The European Commission has not to be part of the agreement because the European Commission is already part of the grant agreement with the coordinator.

Several groups developed *CAMs (CAM)* with the aim to simplify the process of formulating a CA. Consortium agreements models are modular template contracts (EU-Büro des BMBF 2019; Kooperationsstelle EU der Wissenschaftsorganisationen 2019; European IPR Helpdesk 2019).

The most common CAM's are: DESCA, MCARD 2020 IMG4-2020 and EUCAR. The three joint undertakings *Clean Sky 2*, *ECSEL* and *Innovative Medicines Initiative 2* drafted their own CAM's (EU-Büro des BMBF 2019; Kooperationsstelle EU der Wissenschaftsorganisationen 2019; European IPR Helpdesk 2019).

In the current inquiry, we analysed the following five CAM's (listed in alphabetical order):

Clean Sky 2 The *Clean Sky 2 CA* is developed by *Clean Sky* is joint undertaking between the European Commission and the European aeronautics industry. Clean Sky aims to support the European aeronautics research activities to develop more environmentally friendly aircrafts. (Clean Sky 2 2019)

DESCA-2020 *DESCA-2020 (DEvelopment of a Simplified Consortium Agreement for FP7)* is a CA developed by a multinational working group of stakeholders in FP7 namely. DESCA aims to be a simple and comprehensive CA that is "stripped of all unnecessary complexity in both content and language"(DESCA 2019).

EUCAR in Horizon 2020 *European Council for Automotive Research and Development (EUCAR)* developed a CA for the automotive sector (EUCAR 2019).

IMG4-2020 The Aero Space and Defence Industries (ASD) and Industrial Management Group (IMG) drafted a CA that is based on the DESACA model. IMG4-2020 is tailored to European Aeronautics projects and does not contain the options present in the DESCA model. Instead, the IMG4-2020 model includes statements regulating mutual loan of materials (European IPR Helpdesk 2019).

MCARD-2020 Digital Europe an (association representing digitally transforming industries) designed the *Model Consortium Agreement for Research, Development and Innovation Actions under Horizon 2020 (MCARD-2020)* that is adapted to the specific needs of the electronic industry (EPIC 2019).

5.2.2 Description and fixation of the primary and secondary data

5.2.2.1 Primary data

The starting point of our analysis were the *eight* legal documents (see Annex, Table 11). These eight CAMs are representative text documents or text fragments of the *five* CAMs described above. CAM's are modular, formal, legal documents and as such can be considered as noisy text documents in nature. Additionally, those documents are very explicit and repetitive and therefore partially redundant. Hence, we performed a quick perusal to prepare our raw data by filtering out irrelevant parts or parts that could interfere with our main analysis. This has the purpose of anonymising the documents and making the documents easily readable for the coders. Through this initial cleaning process, we improved our analysis significantly.

5.2.2.2 *Secondary data*

We *reduced* the initial *eight* documents to *six* documents, because a short initial review of the documents showed that the DESCA-A (B) and the DESCA-B (C) CAM are largely redundant. Hence, we decided *not* to include the uncommented version DESCA-B in our analysis in order to avoid negative redundancy. For the same reasons, we decided *not* to analyse the (D') EUCAR-B CAM. Hence, we analysed the following six CAM's (in alphabetical ordering):

1. A: Clean Sky 2
2. B: DESCA-2020 with Commentary
3. D: EUCAR in Horizon 2020
4. E: IMG4-2020
5. F: MCARD-2020
6. G: MCARD-2020 Dissemination

That means we have now *six documents* representing *five different CAMs*. If present, we have removed the formalities at the beginning and at the end of each document. This second clean-up process makes the documents more accessible for qualitative and quantitative analysis. For reasons of consistency, we converted all files into PDF. Additionally, to make the files available to our quantitative methods we converted all files into plain text format.

5.2.2.3 *Definition of the units of analysis*

The most ample unit of our *qualitative* content analysis is the text corpus of all the six documents at hand. Words do have semantics but only partially so. Due to known technical limitations and limitations in computer linguistics (text mining), the smallest unit of our *quantitative* analysis is forced to be single characters and words (see previous section: *Quantitative automated topic modelling of open science*). Yet, for our *qualitative* analysis, the smallest unit of analysis coincides with the smallest meaningful, i.e., full semantic unit of text: the full sentence, because we are interested in propositions and propositional attitudes. Coders are advised via the coding book to count full sentences.

Text corpus (see Annex, Table 11)

1. Document count:	6
2. Total pages:	222
3. Total types:	2,000
4. Total tokens:	38,815

5.2.3 Coding book: categories, definitions and coding rules

In our theoretical analysis, we identified 18 important categories of open science. From these categories, we directly derived 18 corresponding *coding book categories* (see Annex, Table 12). We deliberately did not choose technical definitions for the categories, but chose common natural language definitions from the Oxford Dictionary (Lexico.com and Oxford University Press (OUP) 2019) to normalise the fitting of our 18 categories. The coding rules (see Annex, Table 12) were created in such a way that they are as unambiguous and as precise as possible with regard to the definition of the categories.

5.2.3.1 Run coding with two coders

We performed a coding with two independent coders (persons). The two coders did not know the results of each other's categorisation. The coders each received the six documents and the codebook with a short introduction. The coders were encouraged to systematically go through the six documents and follow precisely the coding rules given in the codebook. Based on the coding rules in the coding book (see *Annex*, Table 12), both coders carried out the assignment of text passages to the previously determined categories independently of each other. The results of the two coders can be seen in Table 1 and 2.

Table 1: Coder 1 results

Coder 1																		
Category	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**
Observed	5	4	34	143	8	47	93	8	3	18	2	0	18	55	5	39	1	54
Add 1	6	5	35	144	9	48	94	9	4	19	3	1	19	56	6	40	2	55

Table 2: Coder 2 results

Coder 2																		
Category	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**
Observed	1	7	5	10	8	5	11	7	8	14	1	13	7	10	4	16	0	15
Add 1	2	8	6	11	9	6	12	8	9	15	2	14	8	11	5	17	1	16

5.3 Quantitative content analysis

In order to determine the theoretical probability of the occurrence (the expected frequencies) of each of the previously determined 18 categories within the six documents (our text corpus), it was necessary to perform the following four steps.

In the first step, we created a list of synonyms (see Annex, Table 13) for each of the 18 categories (see Annex, Table 12). The source for the synonyms was the online *Oxford Dictionary* (Lexico.com and Oxford University Press (OUP) 2019). The background assumption here is that the terms of the compiled synonym list sufficiently approximate the respective category. It is important to note, that in this respect, we are not resorting to the necessary or sufficient conditions of a category, but only to a loose, associative connection of terms, that may give rise to a category.

In the second step, we calculated the probability of each term in the list of synonyms by utilising a comprehensive frequency list. As an approximation of the a priori general term frequency, we have used the *Word frequency database* (Word frequency data 2019) containing 450,000,000 types. This procedure enables us to estimate a baseline of expected term frequencies against which we can measure the observed term frequencies within our text corpus (our 6 documents). *We calculated the a priori general category probability as the sum of the individual frequencies of all the synonym types for a corresponding category divided by the total number of types in the general frequency list.* The derived category probabilities can be seen in Annex, Table 13. In this way, we have arrived at a simple model for our categories.

In the third step, we searched our entire text corpus for each synonym term (and its root) for each category and counted the total hits. Our corpus contains 2,000 types in total. Finally, the expected category frequency is equal to the general probability of the previously computed corresponding category times the total number of types in our text corpus. The expected category frequency can be compared with the de facto observed category frequency within our text corpus. The expected and observed category frequencies can be seen in Annex, Table 13. For better readability, the decimal points have been truncated to one decimal digit.

Table 3: Category frequency text search results (cut to one digit after decimal point)

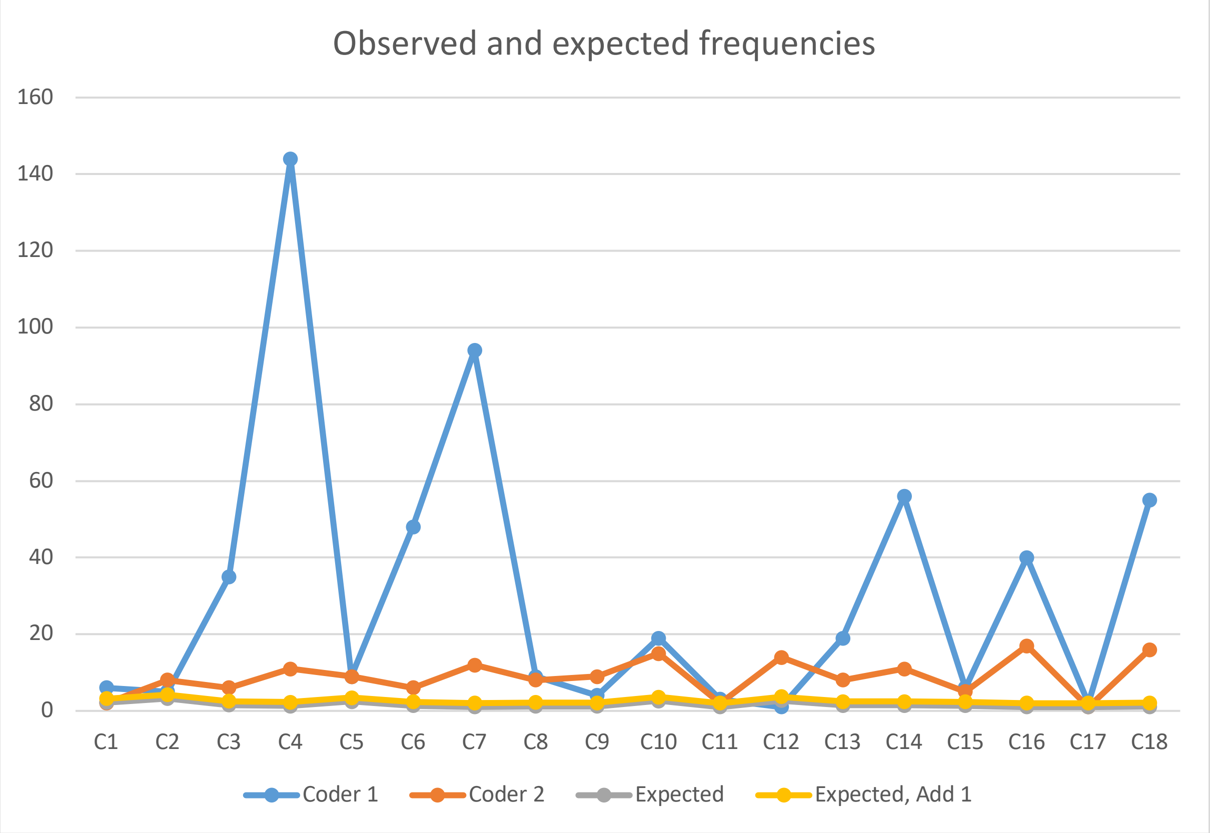
	Text search category frequency																	
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
Observed	61	261	532	381	30	158	57	19	11	306	0	231	46	175	68	100	7	111
Expected (add 1)	2,1	3,2	1,4	1,2	2,4	1,3	1,0	1,2	1,1	2,6	1,0	2,6	1,4	1,4	1,3	1,0	1,0	1,1

These three steps enable us to perform the statistical analysis as shown in the next section: *Statistical analysis*.

5.4 Statistical analysis

For our statistical analysis, we used the free statistical computing software *R* (R 2019). We exported our MS-Excel-sheets to comma separated value files (CSV) and imported the CSV data into R. In particular, we used the R packages *irr* and *ggpubr* (CRAN 2019).

Figure 6: Observed and expected category frequencies



5.4.1 Inter rater reliability

The inter rater (coder) reliability is a statistical measure of accordance between two independent evaluators in general calculated by the relative fit of individual judgments. We performed two different tests: Krippendorff's α (alpha) (Krippendorff 2011), and Kendall rank correlation coefficient τ (tau) (Abdi 2007). The results can be seen in Table 4 and 5. Please see also Figure 6.

To estimate the inter rater (coder) reliability between Coder 1 and Coder 2 we calculated Krippendorff's α (Krippendorff 2011). The result of the test is α = 0.422. Krippendorff's α is a (rather complex) ratio of observed and expected disagreement (Krippendorff 2011). Krippendorff's α can be regarded as a relatively conservative reliability measure. The domain of Krippendorf's α is: 1 ≥ α ≥ 0, where 0 expresses perfect disagreement and 1 expresses perfect agreement between the coders. (Krippendorff 2011) In general, conclusions can be drawn for values α ≥ .8. Yet, tentative conclusions can be drawn for α ≥ .667 (Krippendorff 2004, 241–42). The results can be seen in Table 4. Please see also the next section: *Interpretation of results.*

Table 4: Inter coder reliability A

Krippendorff's α	
Coders	2
Cases	18
Decisions	36
Krippendorff's α (ordinal)	**α = 0.422**

Although we did *not* achieve the level of agreement of α ≥ .667, there *is* however, a *systematic agreement between the two coders*, as shown by the Kendall rank correlation coefficient τ. The result of this test is τ = 0.3446261. The Kendall rank correlation coefficient "evaluates the degree of similarity between two sets of ranks given to a same set of objects" (Abdi 2007). If τ is equal to -1, then there is total negative correlation, if τ is equal to 1, then there is total positive correlation, if τ is equal to 0 there is no correlation (Wikipedia 2020b). We can see that there is a correlation between Coder 1 and Coder 2 (p > .001). The results can be seen in Table 5. Please see also the next section: *Interpretation of results* for a possible explanation of the divergence between the coders.

Table 5: Inter coder reliability B

Kendall rank correlation coefficient τ	
Coders	2
Cases	18
Decisions	36
Z-Score	z = 1.9465
P-value	p = 0.05159
Kendall rank correlation coefficient τ	**τ = 0.3446261**

5.4.2 Testing Coder 1

We performed Fisher’s exact test to measure the significance of the observed category frequencies by coder 1. The observed category frequencies by coder 1 are significant (p < .001). The results can be seen in Table 6. Please see also Figure 6.

Table 6: Coder 1 significance

Fisher's exact test	
Significance level	0.05
Alternative hypothesis	two-sided
P-value	**p = 0.0004998**

5.4.3 Testing Coder 2

We performed Fisher’s exact test to measure the significance of the observed category frequencies by coder 2. The observed category frequencies by coder 2 are *not* significant (p > .6). The results can be seen in Table 7. Please see also Figure 6.

Table 7: Coder 2 significance

Fisher's exact test	
Significance level	0.05
Alternative hypothesis	two-sided
P-value	**p = 0.6567**

5.4.4 Category search frequency

With the help of our category model (synonym lists) (see **Annex**, Table 13), we are able to test our hypothesis. To analyse our contingency table: expected category frequency and observed category frequency, we performed Fisher's exact test (Wikipedia 2019). We performed Fisher’s exact test in favour of the Chi-Square test because our sample size is small and our expected values are small. The category frequency is significantly higher than expected (p > .001). The results can be seen in Table 8. Please see also Figure 6.

Table 8: Category synonyms significance

Fisher's exact test	
Significance level	0.05
Alternative hypothesis	two-sided
P-value	**p = 0.0004998**

5.4.5 Frequent categories

The following *9 categories* are (a) more frequent than expected and (b) have relatively high inter-coder agreement (Table 9, see also Annex, Table 12). Please see also Figure 6.

Table 9: Frequent categories

	Frequent categories
#	Category
C4	Closed / Non-Disclosure / Confidentiality / Privacy / Restrictions / Limits
C5	Public / Society / Community
C7	Patent / Intellectual Property
C8	Knowledge / Knowledge transfer
C10	Data
C13	Quality / Interoperability / Standards / Practices / Best practices / Sustainability / Re-use / Transparency / Verifiability / Falsifiability / Visibility
C14	Ethics / Fairness / Equality / Responsibility
C16	Copyright / Licensing
C18	Software / Source Code

The remaining 9 categories: C1, C2, C3, C6, C9, C11, C12, C15 and C17 (see Annex, Table 12) are *not* significantly more frequent or where *not* consistently measured by the coders. Please see also Figure 6.

5.4.6 Prominent categories

The 5 most prominent categories present in the CAMs among the 9 aforementioned categories are: C7, C10, C14, C16 and C18. We deliberately omit category C16 here because the documents analysed are legal texts and because it overlaps thematically with category C10. (see Table 10, see also Annex, Table 12). Please see also Figure 6.

Table 10: Prominent categories

	Prominent categories
#	Category
C7	Intellectual property
C10	Open source software
C14	Open data
C16	Copyright, Licensing
C18	Ethics and responsibility

6 Interpretation of results

Since the analysed documents are legal documents, it is not surprising that the categories *Copyright, Intellectual Property* appear in the documents. However, it is interesting to note that *digitalisation* does not seem to be particularly relevant in the given CAMs. The low level of relevance of the category of digitalisation could be explained in three ways. First, it could be a systematic error in our analysis. Secondly, digitalisation could be so self-evident among the authors or stakeholders that it is not explicitly mentioned. Thirdly, it could be a blind spot among the authors or stakeholders.

We measured a slight agreement between the coders with Krippendorff's α = 0.422 (see Table 4). The relatively low Krippendorff alpha value can be explained by the fact that (a) both coders did *not* have any former training on the coding method and (b) by the noisiness of the documents due to their formal (legal) nature. Furthermore, a low Krippendorff alpha value can also be interpreted as the strong independence of judgement of the coders. Yet, we measured a systematic agreement between the coders with Kendall rank correlation coefficient ($\tau = 0.3446261$, $p > .001$) (see Table 5). We can observe a systematic agreement of the coders as well as a simultaneous independence of the coders. Therefore, in conjunction, there is indeed evidence for a systematic agreement between the coders. This can be interpreted as evidence in favour of our hypothesis H1.

According to Fisher's exact test (see Table 8), the search category frequency is significantly higher than expected ($p > .001$) by our conservative category model. We measured *no* significantly high category frequency for coder 2 (see Table 7). Yet, we in fact measured significantly high category frequency for coder 1 (see Table 6). In combination with the observed agreement and independence of the coders, this means that it is unlikely that the observed occurrence frequency of the categories relative to expected occurrence frequency of the categories (modelled via our synonym list) is *not* due to chance.

We observed 9 categories, i.e., 50% of the categories to be significantly and consistently more frequent than expected (see Table 9). With our conditional working definition **D** and per modus ponens we can reject the null hypothesis **H0** and have good reasons to believe that *our alternative hypothesis **H1** is true*:

H1: Open science *is* implicitly relevant in the CAMs.

Our results should be treated with caution because our analysis is conducted on a relatively small dataset and we have worked with only two coders, i.e., with a small sample size. Nonetheless, our analysis provides at the very least some initial evidences that *open science is in fact implicitly relevant in the consortium agreement models (CAMs).*

7 Conclusions and implications

7.1 Main conclusion

The main motivation of the present analysis was to get a clear understanding of the current state of open science in the European aeronautics and air transport (AAT) research landscape especially of the content of the consortium agreement models (CAMs).

Open Science is not a concept on which beneficiaries have to commit under the rules for participation. To our knowledge, the H2020 rules for participation does not address the concept of open science but only the sub-concept of open access. To our knowledge, same holds for the MGA and the AMGA. In view of this, it is no surprise that there is no explicit mention of open science in the CAMs.

Yet, our analysis of the five major CAMs that are widely used in the European AAT research landscape shows that *open science and its underlying conceptual framework is indeed relevant in these CAMs*. Because open science is relevant in the analysed CAMs, our strategy for harmonising, the forthcoming open science code of conduct with these CAMs can exploit the fact that some of the underlying open science concepts are already relevant within those CAMs. In particular, we should focus on the most relevant four identified categories (C7) **intellectual property,** (C10) **open data,** (C14) **ethics and responsibility** and (C18) **open source software**. We deliberately omit category *(C16) Copyright and Licensing* here because the documents analysed are legal texts and because it overlaps thematically with category C10. (See Table 9, 10 and Annex, Table 12.) Interestingly one of the key categories of open science and the European AAT research namely **digitalisation** seems to be underrepresented in the analysed documents.

These results are the basis for further development of our implementation strategies for open science in the challenging field of the European AAT research landscape. To develop implementation strategies, communication strategies and argumentation lines for implementing open science into the European AAT research it is important to know the relevance or irrelevance of the discussed concepts to the main stakeholders and their legal instruments. Our analysis shed some light on the current state of open science in the European AAT landscape. Particularly, the current analysis helps the OSCAR project to get a better understanding of the upcoming possible implementations paths.

7.2 Implementation paths

7.2.1 Intellectual property management

The analysed CAMs are legal document templates designed by stakeholders and lawyers in the respective fields of research and development. The purpose of these CAMs is to give consortia of AAT research projects within the EU an easy to use and easy to understand contract template for their projects. The focus of these template documents is naturally on the mutual relationship between the project partners, their rights, and obligations. These contracts also regulate how the work and the results are distributed.

In particular, non-disclosure agreements are made in these contracts. In general, intellectual property (IP) management is practised. In this respect, it is no surprise that IP is a significant category in the CAMs analysed. Open science and IP *prima facie* seem to contradict each other. However, there is a tension-rich and at the same time fertile interface between open science and IP as well as IP management. In the further course of the project, we should address this challenging interplay between open science and IP in more detail.

In the OSCAR project, we should not try to change existing patterns of contract formation or the contracts themselves, but we should focus on the development of reasonable opt-in opt-out models that emphasis the compatibility of standard contracts and open science. These harmonised opt-in, opt-out models should be made easily and obviously available to consortium members and stakeholders from the very beginning of the respective project.

7.2.2 Ethics and responsibility

One of the important categories of open science we found to be implicitly relevant in the CAMs is ethics and responsibility. This category is not only an important category of open science but it is also an important category for the European science and research landscape in general. Ethics and responsibility is closely intertwined with responsible research and innovation (RRI): it is about involving society and its values in the scientific process. (European Commission 2020) RRI is about ethics, public engagement, gender equality, governance, open access and science education (European Commission 2020). RRI is "[a] cross-cutting issue in Horizon 2020" (European Commission 2020) and it is a "key action of the 'Science with and for Society'" (European Commission 2020).

In the OSCAR project, we should emphasis on the existing guidelines and best practices of RRI already established by the European research landscape.

7.2.3 Open data and open source

Open data and open source software are part of open science. Open data and open source are fundamentally linked to other categories of open science like open access. We found that open data and open source are implicitly relevant categories in the CAMs. In the further course of the OSCAR project, we should utilise the fact that the stakeholders already know those categories. We could base our argumentation lines on the fact that there are great, successful projects already using best practises and standards of open source software and open data. In this way, we can close the gap between the status quo of European AAT projects and the next generation of European AAT projects.

7.2.4 Digitalisation

On interesting result of our analysis is that the category of digitalisation seems to be not relevant in the given CAMs. This is strange because digitisation is without doubt one of the biggest global trends. As already discussed in section: *Interpretation of results* the absence of digitalisation in the CAMs may be due to a blind spot of the stakeholders or the obviousness of the subject. Yet, digitalisation is one of the key drivers for open science; it enables many principles and paradigms of open science to be feasible in the first place. On the one hand, digitalisation is key driver for open science; on the other hand open science enforces digitalisation. We should focus on the mutual reinforcement of open science and digitalisation as well as on the fact that digitalisation is an inevitable necessity for all projects.

7.2.5 Communication strategy and incentives

In the further course of the OSCAR project, we should focus on developing communication strategies in accordance with the results of the OSCAR project so far. Our communication strategy specifically should address (a) the main stakeholders of AAT and (b) the European Commission.

One of the most important pivotal points for the successful implementation of open science is the creation incentive structures. Such incentive structures cannot be generated by the scientific community alone, but also require clear incentives and commitments on the part of policy makers. Eva Méndez, chairwoman of the Open Science Policy Platform (OSPP) (OSPP 2020), gets to the very core of this topic: to implement open science in the European research landscape we need a three level approach. (Méndez 2019) To make open science happen we need the following three actions (Méndez 2019):

1. **Create incentives**: Scientists need suitable incentive structures that motivate them to be more open with their research.
2. **Establish clear rules**: Scientists need clear rules that guides them in the active effort to do open science.
3. **Provide guidance**: Scientists need to be taught how to open up their research according to the principles of open science.

The same holds true for implementing open science in the European AAT research landscape. Our communication strategy should focus not only on the identified key concepts: intellectual property, open data, ethics and responsibility and open source software but also on those three key strategical actions to make open science happen.

8 References

Abdi, Hervé. 2007. 'The Kendall Rank Correlation Coefficient'. In *Encyclopedia of Measurement and Statistics*, edited by Neil Salkind. Thousand Oaks (CA): Sage.

Bezjak, Sonja, April Clyburne-Sherin, Philipp Conzett, Pedro Fernandes, Edit Görögh, Kerstin Helbig, Bianca Kramer, et al. 2018. *Open Science Training Handbook*. Zenodo. https://doi.org/10.5281/ZENODO.1212496.

Bhattacherjee, Anol. 2020. *Social Science Research: Principles, Methods, and Practices*. Open Textbook Library. https://open.umn.edu/opentextbooks/textbooks/social-science-research-principles-methods-and-practices.

Blasius, Jörg, and Nina Baur. 2014. 'Multivariate Datenanalyse'. In *Handbuch Methoden der empirischen Sozialforschung*, edited by Nina Baur and Jörg Blasius, 997–1016. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-18939-0_79.

Clean Sky 2. 2019. 'Welcome to the Clean Sky | Clean Sky'. 2019. https://www.cleansky.eu/.

CRAN. 2019. 'The Comprehensive R Archive Network'. 2019. https://ftp.fau.de/cran/.

DESCA. 2019. 'DESCA | DESCA 2020 Model Consortium Agreement'. 2019. http://www.desca-2020.eu/.

EPIC. 2019. 'DIGITALEUROPE MCARD-2020 Version 2.2'. *DIGITALEUROPE* (blog). 2019. https://www.digitaleurope.org/resources/digitaleurope-mcard-2020-version-2-2/.

EU-Büro des BMBF. 2019. 'Konsortialvertrag - Horizont 2020'. 2019. https://www.horizont2020.de/projekt-konsortialvertrag.htm.

EUCAR. 2019. 'Horizon 2020'. *EUCAR* (blog). 2019. https://eucar.be/horizon2020/.

European Commission. 2020. 'About RRI - RRI Tools'. 2020. https://www.rri-tools.eu/about-rri.

———. 2020. 'Responsible Research & Innovation'. Text. Horizon 2020 - European Commission. 2020. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation.

European IPR Helpdesk. 2019. 'European IPR Helpdesk: Get Your Ticket to Innovation! | European IP Helpdesk'. 2019. https://www.iprhelpdesk.eu/.

Fecher, Benedikt, and Sascha Friesike. 2014. 'Open Science: One Term, Five Schools of Thought'. In *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing*, edited by Sönke Bartling and Sascha Friesike, 17–47. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2.

Jurafsky, Daniel, and James H. Martin. 2019. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, N.J: Pearson Prentice Hall.

Kooperationsstelle EU der Wissenschaftsorganisationen. 2019. 'KoWi - Model Consortium Agreements'. 2019. https://www.kowi.de/en/kowi/proposal-project/contract-management/consortium/model-consortial-agreement/model-consortium-agreements.aspx.

Krippendorff, Klaus. 2011. 'Computing Krippendorff's Alpha-Reliability'. *Departmental Papers (ASC)*, January. https://repository.upenn.edu/asc_papers/43.

Lexico.com, and Oxford University Press (OUP). 2019. 'English | Lexico'. Lexico Dictionaries | English. 2019. https://www.lexico.com/en/english.

Margolis, Eric, and Stephen Laurence. 2019. ‘Concepts’. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/concepts/.

Mayring, Philipp. 2014. *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution*. Klagenfurt, Austria: SSOR.

Méndez, Eva. 2019. ‘“Open Science?… Darling, We Need to Talk!” | EInfra Central’. 2019. https://einfracentral.eu/news/open-science-darling-we-need-talk.

OSPP. 2020. ‘Open Science Policy Platform | Open Science - Research and Innovation - European Commission’. 2020. https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-policy-platform.

R. 2019. ‘R: The R Project for Statistical Computing’. 2019. https://www.r-project.org/.

Wikipedia. 2019. ‘Fisher’s Exact Test’. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Fisher%27s_exact_test&oldid=933086673.

———. 2020a. ‘Open Science’. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Open_science&oldid=937016095.

———. 2020b. ‘Kendall Rank Correlation Coefficient’. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Kendall_rank_correlation_coefficient&oldid=938555917.

———. 2020c. ‘Topic Model’. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Topic_model&oldid=945377340.

Word frequency data. 2019. ‘Word Frequency: Based on 450 Million Word COCA Corpus’. 2019. https://www.wordfrequency.info/.

9 Appendix

All data, statistics and media used or generated by this analysis can be found in the zip files: "data.zip", "stats.zip" and "media.zip".

Table 11: Document mapping

Int. ID	Short Title	Original Title, Description	Authors	Date	Version	File Format	Pages	Types (Words)	Tokens	Characters	Analysed
A	Clean Sky	"Clean Sky 2 Joint Undertaking, Consortium Agreement"	Clean Sky 2 Joint Undertaking	21.12.2017	-	.pdf	58	1179	9288	107778	Yes
B	DESCA-A	"DESCAR Horizon 2020 Model Consortium Agreement", "Elucidations & Comments"	ANRT, EARTO, KoWi, LERU, VTT, ZENIT (coordinated by Fraunhofer and Helmholtz Association)	2016-03	Version 1.2	.pdf	54	1344	11229	131606	Yes
C	DESCA-B	"DESCAR Horizon 2020 Model Consortium Agreement"	ANRT, EARTO, KoWi, LERU, VTT, ZENIT (coordinated by Fraunhofer and Helmholtz Association)	2017-10	Version 1.2.4	.docx	-	-	-	-	No
D	EUCAR-A	"EUCAR Model Consortium Agreement for the Horizon 2020 Framework Programme for Research and Innovation"	EUCAR	11.06.2014	V.5	.docx	25	745	3179	35536	Yes
D'	EUCAR-B	"EUCAR Model Consortium Agreement Horizon 2020"	EUCAR	2018-03	-	.docx	-	-	-	-	No
E	IMG4	"IMG4 - H2020 H2020 FRAMEWORK PROGRAMME"	ASD, IMG	08.07.2014	Final	.docx	41	953	5767	67718	Yes

Int. ID	Short Title	Original Title, Description	Authors	Date	Version	File Format	Pages	Types (Words)	Tokens	Characters	Analysed
F	MCARD-A	“MODEL CONSORTIUM AGREEMENT FOR RESEARCH, DEVELOPMENT AND INNOVATION ACTIONS UNDER HORIZON 2020”	DigitalEurope	14.12.2017	V.2.0	.docx	39	1182	8318	99900	Yes
G	MCARD-B	“Decisions upon proposals for the plan for use and the Dissemination of Results”	DigitalEurope	14.12.2017	V.2.0	.docx	5	316	1034	12550	Yes

Table 12: Coding book

#	Category	Definition (from the Oxford Dictionary)	Coding Rule	Example	Reference
C1	Openness	Lack of restriction; accessibility	1. If the authors write about openness in general. 2. If the authors write about the absence of restrictions or limitations in general.	"It makes access "Needed for the Project" very open in order to make work on the Project as uncomplicated as possible."	DESCA-2020_2016-03, 7
C2	Sharing / Giving / Contribution / Collaboration	1. Have a portion of (something) with another or others; 2. Give (something, especially money) in order to help achieve or provide something 3. The action of working with someone to produce something.	1. If the authors write about sharing with others or about giving. 2. If the authors write about rely on the help of someone or something. 3. If the authors write about provide something to others. 4. If the authors write about collaboration in general.	"Prior written notice of the final version of any planned publication shall be given to the other Parties at least forty-five (45) days before the planned publication submission date."	MCARD-2020_2017-12-14, 19
C3	Accessibility / Availability	1. Something to be able to be reached or entered; 2. Able to be used or obtained; at someone's disposal	1. If the authors write about that something is or is not/should or should not be reachable, accessible or available. 2. If the authors write about something to be obtained or obtainable.	"The wording aims to be accessible and easy to understand notably for non-lawyers."	DESCA-2020_2017-10, 3
C4	Closed / Non-Disclosure / Confidentiality / Privacy / Restrictions / Limits	1. Not open; beeing locked or sealed; 2. The action of making new or secret information known; 3. Belonging to or for the use of one particular person or group of people only; 4. A limiting condition or measure, especially a legal ones	1. If the authors write about something is or is not/should or should not be open. 2. If the authors write about secret or confidential knowledge or information etc. 3. When the authors talk about something being private or belonging exclusively to a group or company or person. 4. If the authors write about limiting conditions or restrictions in general.	"Allowing to produce research results which are available to the third party but which contain hermetically-sealed Results from the Project"	DESCA-2020_2016-03, 31

#	Category	Definition (from the Oxford Dictionary)	Coding Rule	Example	Reference
C5	Public / Society / Community	1. Concerning the people as a whole; a singular or plural; Ordinary people in general; 2. Group of people sharing or having something in common.	If the authors write about the public, the society or the community.	"own internal research and collaborative research (with or without public funding)."	DESCA-2020_2016-03, 25
C6	Publication / Dissemination / Distribution / Deployment	1. Give a share or a unit of (something) to each of a number of recipients; 2. The action or fact of spreading something, especially information, widely	1. If the authors write about to deploy, share or give something. 2. If the authors write about spreading or deploying something. 3. If the authors write about to publish something.	"To ensure that internal distribution of Confidential Information"	Clean_Sky_2017-12-21, 27
C7	Patent / Intellectual Property	1. A government authority or licence conferring a right or title for a set period, especially the sole right to exclude others from making, using, or selling an invention; 2. Intangible property that is the result of creativity, such as patents, copyrights, etc.	1. If the authors write about patents. 2. If the authors write about intellectual property or ownership in general.	"Each Party warranties that to the best of its knowledge, the intellectual property rights it provides as Background and Results"	Clean_Sky_2017-12-21, 11
C8	Knowledge / Knowledge transfer	1. Facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject 2. Move or copy information from one medium, device or context to another;	If the authors write about knowledge, skills, experience, education, information or understanding.	"intangible output of the Action, such as data, knowledge and information"	MCARD-2020_2017-12-14, 6
C9	Value / Added value	The regard that something is held to deserve; the importance, worth, or usefulness of something	If the authors write about worthiness, potential, importance, value proposition, etc.	"for example the actual or potential value"	EUCAR-2020_2014-06-11, 4

#	Category	Definition (from the Oxford Dictionary)	Coding Rule	Example	Reference
C10	Data	Facts and statistics collected together for reference or analysis	If the authors write about data, statistics, analysis, data science, measurement series etc.	“exchange of Project related data and deliverables”	Clean_Sky_2017-12-21, 39
C11	Repository	A place where or receptacle in which things are or may be stored	If the authors write about data or software storage.		
C12	Governance	The action or manner of governing a state, organization, etc.	If the authors write about authority, guidance, state of affairs, conducting a policy etc.	“accordance with the governance structure of the Project, any significant information”	IMG4-2020_2014-07-08, 7
C13	Quality / Interoperabili ty / Standards / Practices / Best practices / Sustainability / Re-use / Transparency / Verifiability / Falsifiability / Visibility	1. The standard of something as measured against other things of a similar kind; the degree of excellence of something; 2. For computer systems or software to be able to exchange and make use of information and data; 3. The actual application or use of an idea, belief, or method, as opposed to theories relating to it; 4. The ability to be maintained at a certain rate or level; Use again or more than once 5. Able to be checked or demonstrated to be true, accurate, or justified; 6. The condition of being transparent; 7. The state of being able to see or be seen	1. If the authors write about excellence, quality or standards in general. 2. If the authors write about the (possibility of) exchange of information. 3. If the authors write about good or bad or best practices, methods etc. 4. If the authors write about sustainability or maintenance.; 5. If the authors write about use, re-use or of something. 6. If the authors write about to justify, verify, falsify or proof something. 6. If the authors write about transparency in general. 7. If the authors write about visibility in general.	“Poor quality of work or reports may be considered to be a breach.”	DESCA-2020_2016-03, 25

#	Category	Definition (from the Oxford Dictionary)	Coding Rule	Example	Reference
C14	Ethics / Fairness / Equality / Responsibility	1. Impartial and just treatment or behaviour without favouritism or discrimination; 2. The state of being equal, especially in status, rights, or opportunities; 3. The state or fact of having a duty to deal with something or of having control over someone	1. If the authors write about equality, fairness, discrimination, etc. 2. If the authors write about responsibility, duty or control.	"Such Access Rights shall be granted on fair and reasonable conditions"	IMG4-2020_2014-07-08, 24
C15	Infrastructure / Platform	The basic physical and organizational structures and facilities (e.g. buildings, roads, power supplies) needed for the operation of a society or enterprise	If the authors write about organisational or technological structures.	"right to receive source code or object code ported to a certain hardware platform"	Clean_Sky_2017-12-21, 27
C16	Copyright / Licensing	1. The exclusive and assignable legal right, given to the originator for a fixed number of years, to print, publish, perform, film, or record literary, artistic, musical or scientific material; 2. Authorize the use, performance, or release of (something)	1. If the authors write about copyright; 2. If the authors write about licensing.	"without implying or granting any license under any patent and copyright of the Disclosing Party"	MCARD-2020_2017-12-14, 30
C17	Digitalisation	The conversion of text, pictures, or sound into a digital form that can be processed by a computer	If the authors write about the transition or the process of digitalisation, conversion of analog material into digital representations, or about the state of beeing digital etc.	"businesses and citizens to benefit fully from digital technologies and for Europe to grow"	MCARD-2020_2017-12-14, 39
C18	Software / Source Code	The programs and other operating information used by a computer.	If the authors write about software, programs, algorithms, etc.	"granting of Access Rights (e.g. the use of open source code software in the Project)"	IMG4-2020_2014-07-08, 23

Table 13: Relative category frequencies

Relative category frequencies (Synonyms)									
	C1	C2	C3	C4	C5	C6	C7	C8	C9
Document Probability	0,0305	0,1305	0,266	0,1905	0,015	0,079	0,0285	0,0095	0,0055
General Probability	0,000583576	0,001115538	0,000246944	0,000125078	0,000742387	0,000190284	1,33867E-05	0,000111856	7,70822E-05
Expected Frequency	1,167151111	2,231075556	0,493888889	0,250155556	1,484773333	0,380568889	0,026773333	0,223711111	0,154164444
Sum	61	261	532	381	30	158	57	19	11
Average	12,2	43,5	133	47,625	10	26,33333333	28,5	9,5	3,666666667
Minimum	0	0	0	0	0	0	11	0	0
Maximum	48	103	465	202	22	58	46	19	10

Relative category frequencies (Synonyms)									
	C10	C11	C12	C13	C14	C15	C16	C17	C18
Document Probability	0,153	0	0,1155	0,023	0,0875	0,034	0,05	0,0035	0,0555
General Probability	0,000818502	2,42067E-05	0,000833589	0,000238753	0,000228727	0,000184371	2,43178E-05	0,00003344	5,32311E-05
Expected Frequency	1,637004444	0,048413333	1,667177778	0,477506667	0,457453333	0,368742222	0,048635556	0,06688	0,106462222
Sum	306	0	231	46	175	68	100	7	111
Average	102	0	38,5	3,833333333	35	13,6	50	7	55,5
Minimum	18	0	4	0	2	0	6	7	34
Maximum	267	0	113	18	87	41	94	7	77