**Deliverable 1.1:** User requirements and specification of the use cases

Irina Bienia, Angela Fessl, Franziska Günther, Sabrina Herbst, Annalouise Maas, Michael Wiese

31/03/2017

Work Package 1: Requirements and use case development

**TraininG towards a society of data-saVvy inforMation prOfessionals to enable open leadership INnovation**

| Dissemination level | PU |
|---|---|
| Contractual date of delivery | 31/03/2017 |
| Actual date of delivery | 31/03/2017 |
| Deliverable number | D1.1 |
| Deliverable name | User requirements and specification of the use cases |
| File | MOVING_d1.1_v1.0.doc |
| Nature | Report |
| Status and version | Final v1.0 |
| Number of pages | 122 |
| WP contributing to the deliverable | 1 |
| Task responsible | TUD |
| Other contributors | EY<br>KNOW |
| Author(s) | Sabrina Herbst, Franziska Günther<br>TUD<br><br>Michael Wiese, Annalouise Maas, Irina Bienia<br>EY<br><br>Angela Fessl<br>KNOW |
| Quality Assessors | Chrysa Collyda, Vasileios Mezaris<br>CERTH |
| EC Project Officer | Hinano SPREAFICO |
| Keywords | requirement analysis, use cases, mock-ups |

## Executive Summary

This deliverable "D1.1: User requirements and specification of the use cases" provides the initial requirement analysis based on the project objectives, the inputs collected from the project partners, as well as the end user groups targeted by the project and the specifications of the platform according to the documented requirements. The document is structured into three major parts. Sections 2 and 3 describe the use cases methodology approaches taken by each partner in pursuing the requirement analysis, each with a structured description of the steps followed. The use cases requirement analysis, which leads to the detailed descriptions of these requirements, is described in Sections 4 and 5, and the platform requirements are extensively described in Sections 6 and 7. Finally, we summarise the achievements and give an outlook towards the ongoing tasks (Tasks 1.2, 1.3 and 1.4).

## Table of contents

# 4 Use case requirements use case 1: research on business information by public administrators ........................................................... 46

# 5 Use case requirements use case 2: managing and mining research information................................................................................... 62

## List of Figures

## List of Tables

## Abbreviations

| Abbreviation | Explanation |
|---|---|
| ATS | Adaptive Training Support |
| bwFDM | Baden-württembergisches Forschungsdatenmanagement |
| DoA | Description of action |
| DUNS | Data Universal Numbering System for the unique identification of entities |
| GA | General Assembly |
| GAAIT | Global Accounting and Auditing Tool |
| GSA | Germany, Switzerland, Austria |
| HTML | Hyper Text Markup Language |
| ISA | International Standard(s) on Auditing |
| MOOC | Massive Open Online Course |
| QA | Quality Assessment |
| SDK | Software Development Kit |
| ToC | Table of content |

# 1 Introduction

This section presents the history and the purpose of the document.

## 1.1 History of the document

**Table 1:** History of the document.

| Date | Version |
|------|---------|
| 30th of January 2017 | v.01: initial table of content (ToC) draft |
| 17th of February 2017 | v.02: revised ToC draft and early description of content send to Quality Assessment (QA) |
| 01st of March 2017 | v.03: incorporated input for ToC from QA and General Assembly (GA) |
| 15th of March 2017 | v.04: review-draft ready for QA |
| 28th of March 2017 | v.05: review-revised draft |
| 31st of March 2017 | v1.0 |

## 1.2 Purpose of the document

This document provides an initial set of requirements for the MOVING platform. These requirements are the central point of D1.1 and the background for the upcoming deliverables D1.2 (Initial implementation of the user studies) and D1.3 (Initial evaluation, updated requirements and specifications). These deliverables will further extend the work of D1.1, followed then by D1.4 (Final implementation of user studies and evaluation).

## 1.3 Structure of the document

This document is structured in three major parts. Sections 2 and 3 describe the methodology of collecting and analysing the requirements for each use case. The next part of the document, Sections 4 and 5, explain the requirement analysis results with detailed scenarios and requirements tables. The functional requirements are displayed in Section 6 as tables including all requirements from both use cases. The initial interaction workflows are displayed in Section 7 through mock-ups. The last section summarises the contributions of this deliverable and gives an outlook for future deliverables and tasks.

# 2 Methodology use case 1: research on business information by public administrators

This section describes the methodology Ernst and Young (EY) followed when performing the requirement analysis for the first use case. In subsection 2.1, we integrate the use case 1 approach, as described in the Description of Action (DoA), and including the descriptions of the two scenarios. We continue with information about the two approaches, interviews in subsection 2.2 and brainstorming sessions with the EY Innovation Management team in subsection 2.3; also, the review of existing literature and tools performed and the alignment with the regulatory requirements (especially the International Standards on Auditing) in subsection 2.4.

## 2.1 Use case description of use case 1

The general objective of MOVING is to improve the social capacity for innovation by expanding its competencies in digital information management through training and empowering people from all societal sectors and disciplinary backgrounds. This should be reached by applying data analytics tools and techniques in the daily working routines of the people. To that extent, an open and interdisciplinary platform will come to aid. The use case described here addresses public administrators, and is provided by EY and their 60.000 compliance officers. The more specific objective here is the application of a human-cantered design approach. According to the scenarios described for this use case, we applied, for the requirement analysis, a mixed-methods design conducting a series of internal interviews and brainstorming sessions based on the questionnaire analysis of responses and performing a review of existing text mining tools.

### 2.1.1 Scenario 1: research on compliance to European laws and regulations

Any organisation, research institutes, universities and companies likewise need to be compliant with the increasing number of current laws and regulations. As non-compliance is a risk that may harm a business seriously, risk assessment procedures need to be performed about compliance with future regulations as well. In this scenario, the compliance officer Mr. Clark has been assigned the task to identify potential risks how the economic and financial changes in law and regulations in the European market may affect the organisation's compliance. For example, changes in the administrative procedures of research organisations and industries are required when participating in recent EU-funded programs. Organisations now need to apply the full-cost pricing model when reporting their costs. Another case where organisational changes in the administration are relevant is when the company plans to extend its business to other European countries. Here, the company needs to conduct so-called "country by country reporting", which refers to a requirement of the EU to companies to split up financial reports by the different countries in which they have businesses. Overall, many jurisdictions require research organisations and companies to constantly being updated and trained in risk assessment procedures to ensure compliance with laws and regulations.

The MOVING platform will provide the user with the appropriate tools to achieve that efficiently and effectively. As part of his research, Mr. Clark is using the MOVING platform to conduct an analysis of the political, economic, social and technological factors that has to be performed. This "PEST-Analysis" is usually performed as part of a "SWOT-Analysis", identifying strengths, weaknesses, opportunities and threats. The MOVING platform enables the user to work on that step by step:

- Topic-based filtering: The user selects from pre-defined filters in which areas the search has to be performed. These are political, economic, social and technological. Multiple selections are possible as well as subtopics, e.g. economic could be split into macro-economical, tax-related, business administrative, managerial accounting and compliance topics. Let us say the user decides to concentrate on the macroeconomic category.

- Faceted search: The user may determine to limit the search to certain countries (where the publication has been made) or the language of those publications. Further filtering concerning a date range, length of publications, and type of publication, industry and ratings of other users may apply as well. In the scenario, the user might decide to concentrate on European countries, publications within the last 5 years, professional journals with a minimum length of 3 pages and published in English, French and German.

- As a first view, the user will see a tag cloud presentation of all topics identified by the MOVING platform. The size of the tag is determined by the importance of the topic. This importance may be based on the frequency of words, i.e. number of publications containing these words. At any stage, the user will be able to drill down into the individual publications available as full text. We expect that topics like credit crisis, tax evasion, nonperforming loans, market barriers, sourcing and emerging markets will appear.

- As an alternative analysis instrument, the MOVING platform will provide a text network. The user can also enable semantic analysis of a search result and view it as topic network highlighting the specific economics subjects and their relations. This network view will not only show the importance of individual topics, but also the strength of the connection between different topics. The user may drill down into a topic, which will then show a sub-network on keywords inside a topic.

- At any stage of the analysis, the user may add or remove filters and select/unselect topics on which to concentrate. In this scenario, the user may remove the date filter and may concentrate on the search term "sourcing" (looking for information for assessing certain production factors). The idea might be to determine whether the availability of certain raw materials is a risk when doing a business inside Europe. By using measures (e.g. centrality in the network), the compliance officer may be able to determine whether sourcing in this case is really an issue.

- To be able to perform risk-based prognosis, the compliance officer will need to see the evolution of topics over time. To enable this, the MOVING platform will contain a lifecycle view for a topic selected. With an adjustable period, the measures above will be used to see when a topic appears, how it develops/evolves and maybe degenerates over time. The user in the scenario

may determine that sourcing was an issue in certain areas and industries in the past but is no longer critical to European business locations. For other topics, e.g. tax planning this might be different which is visible as the topic is still evolving.

The advantage of this new MOVING approach is to generate knowledge from the data at the desired level of detail, e.g. without a need to go into too much detail. However, at any stage, the user is able to access the content if needed. Further, measures are implemented, enabling the user to access the completeness of a research. For example, by using measures of centrality of a topic, the user may assess the most important aspects of a topic. The flexibility of the analysis together with different views and data analysis instruments enables to make decisions on a reasonable basis.

### 2.1.2 Scenario 2: innovation in advisory services

The success of professional advisory services is highly dependent on identifying trends and innovations and being trained in transforming those into solutions for different market sectors, industries and clients. With the instruments detailed in the scenario above the user will be able to perform broad and flexible searches for those trends and think "out of the box". The user may pre-select the country and industry the company is operating in and limit the analysis to economic topics. The approach to focus on certain areas will be supported by the instruments described above including the tag cloud and network view on the data. To determine the need to prepare for future regulations, the user may select the general political category and use the lifecycle view to see which topics are evolving. One of those topics may be the "country by country reporting" promoted by the EU. From the individual publications, the user may determine which steps are necessary to prepare for this change in regulations.

## 2.2 Interviews

For the analysis of requirements, we conducted interviews within EY on the topic of "Unstructured Data and Information". Our goal was to get a deeper understanding of how auditors of EY work.

This includes the different working processes, necessary documents and tasks or activities.

Furthermore, we included junior and experienced professionals from other service lines in order to collect extensive insight on their specific needs. Along with the Assurance service line, we also tried to reach the Advisory, TAX and the Transaction Advisory Services, as well as Business Development, IT and Sector leaders. From a practical point of view, our goal was to discover possible entries for the MOVING platform, which focus on implementing the needs of an auditor.

### 2.2.1 Questionnaire

We list below the questions included in the questionnaire:

1) How important is, in your opinion, the role of "Unstructured Data and Information" within the framework of your professional activity?
2) What type of unstructured data do you have to deal with within your professional activity?

3) Which aspects of the data or in what service/business questions are you analysing?

4) What is the volume of data (GB, number of pages, number of lines, etc.)?

5) What do you do when analysing unstructured data? Please, describe your approach.

6) How accessing the data?

7) Do you use digital applications (e.g., desktop-based software or online-based tools) to evaluate these data? If so, which tools? Are there tools that you prefer using?

8) Which tools do you use for which work steps? Please describe.

9) Which devices do you use for which work processes / work steps?

10) Is there a need to automate work steps, entirely or partially? Describe what steps these would be, and how automation might look from your point of view, or how automation would be useful in your view. How would such a scenario look like?

11) Is it necessary to provide specific evidence or proof of evidence? What do they look like?

12) What are the requirements for the documentation?

13) What problems do you have in your daily work with unstructured data (e.g., data volumes, structure, data sources, data access), where do you see improvement opportunities?

14) Are there any specific functionalities that you already consider necessary, e.g. Network images, tag clouds, filter functions, thesauri?

15) Are there specific security requirements (e.g., privacy) for the unstructured information and data you are working with?

16) How do you estimate your knowledge / skills for working with large unstructured data sets? Methodology in the field of text mining? Where do you see further education requirements?

17) In which form have you already participated in further training on text mining, work with unstructured data, etc.?

18) In which form do you continue to inform and educate yourself? Which further formats do you use, which formats do you prefer?

19) What experiences have you had with digital formats in company / non-company training, which of the following formats are familiar to you, what have you already tried: Blended Learning, Game Based Learning and Massive Open Online Courses (MOOCs) Web-based training.

20) What other, above-mentioned digital formats in the field of company / non-company training have you already had experience with?

### 2.2.2   Analysis of responses

The questionnaire and the analysis of responses are key parts in the mixed-methods approach we used during the requirements analysis. For the response analysis, we present here the responses of 26 junior and experienced auditing professionals who took part in the survey, analysed and interpreted. Almost all responses of the 20 questions of the questionnaire were assessable. The answers from questions 3, 8, 10, 11, 13 and 19 vary too much; hence, we cannot summarise and visualise them. The importance of unstructured data and information within the framework of an auditor's professional activity is the focus of the questionnaire (Figure 1). Almost 90% of the

interviewees recognise the importance of unstructured data and the high priority within the future work framework. In their everyday routine, they deal with a variety of unstructured data (Figure 2), ranging from contracts to internet sources and from databases to work-specific sources of data such as invoices, delivery notes and EY internal tools like Global Accounting and Auditing Information Tool (GAAIT[1]).
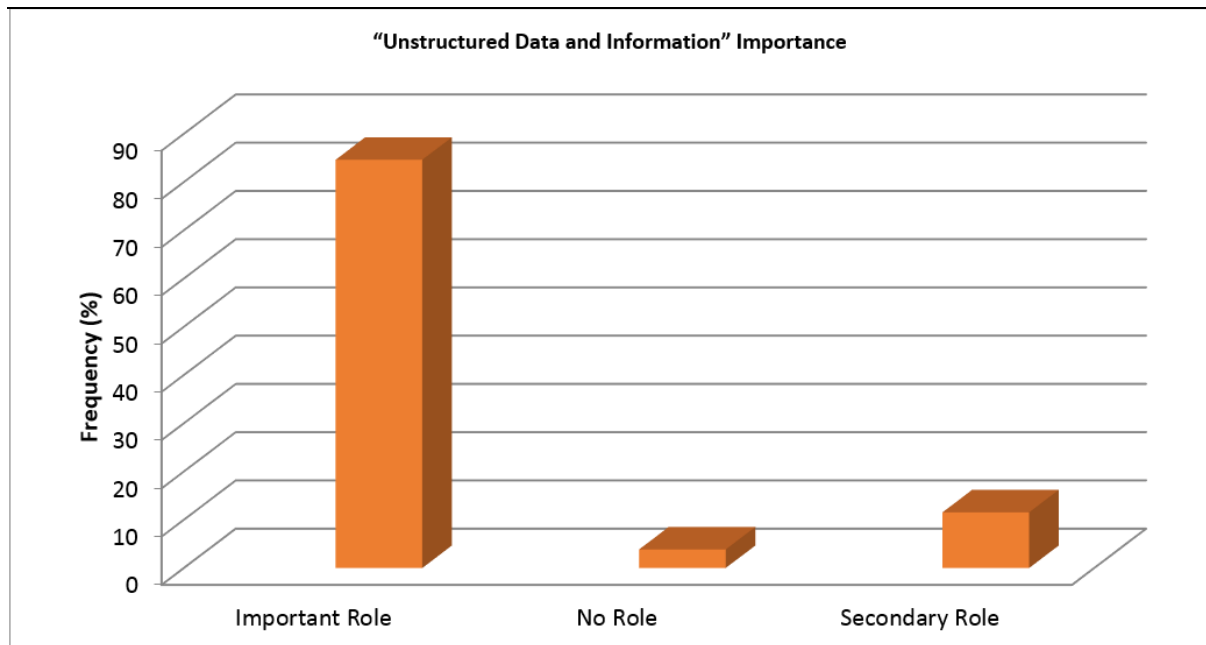


**Figure 1:** 1st question ("How important is, in your opinion, the role of "Unstructured Data and Information" within the framework of your professional activity?").

When asked, what approach they take when analysing the unstructured data they encounter, in their everyday professional activities, the interviewees gave multiple examples amongst which search terms, critical screening and summaries/overviews were most often mentioned (Figure 3).

[1] http://www.ey.com/gl/en/services/assurance/assurance-key-a-a-guidance-on-ernst---young-online---global-accounting---auditing-information-tool (2017-03-27)
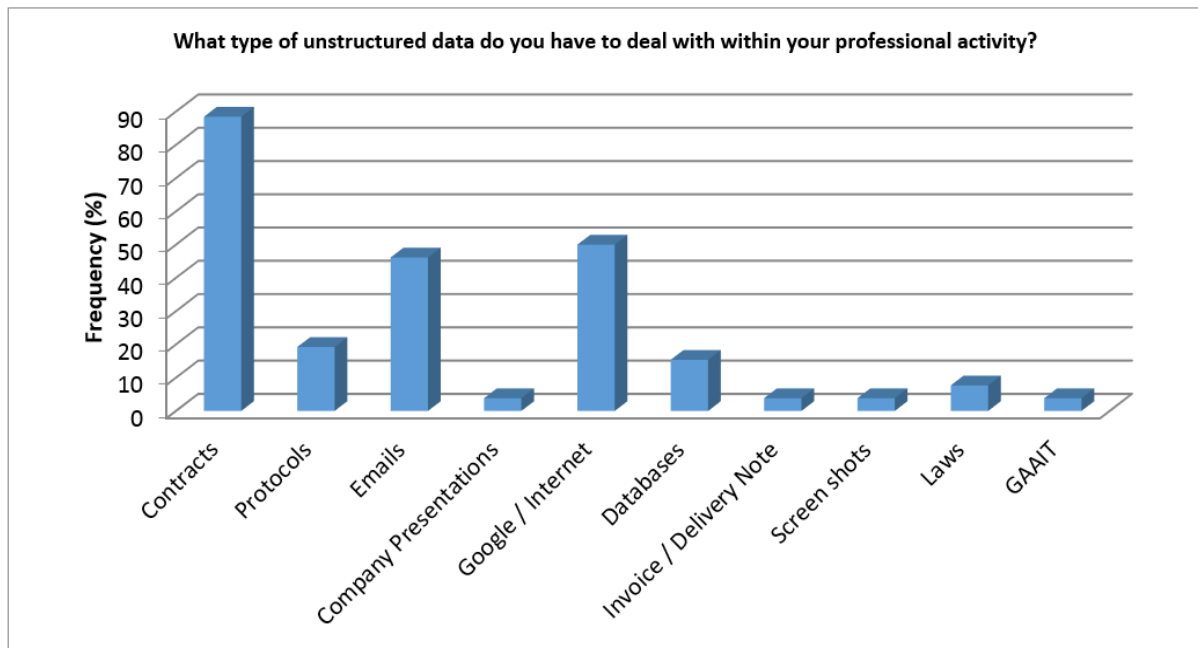
**Figure 2:** 2nd question ("What type of unstructured data do you deal with within your professional activity?").
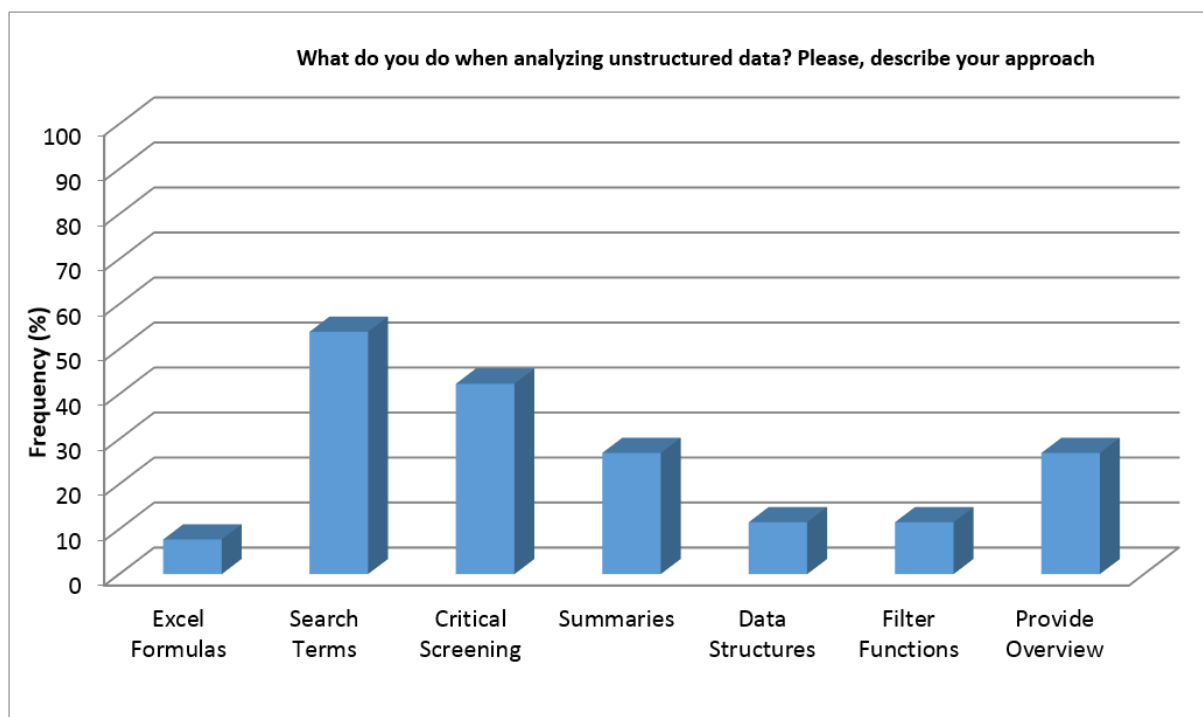
**Figure 3:** 5th question ("What do you do when analysing unstructured data? Please, describe your approach.").

Moreover, the interviewees pointed to specific properties unstructured data should have in order to facilitate the use of such data. Out of the five properties mentioned, comprehensibility is most relevant, while structure, referencing, conciseness and relevance sum up to less than comprehensibility (Figure 4). However, almost half of the interviewees did not answer this question, which creates a difficulty throughout the assessment of responses. For the continuous training assessment among auditors, we asked the interviewees to describe self-teaching methods and channels they use for self-improvement throughout their professional activities. The responses

(Figure 5) are detailed and diverse, ranging from office/classroom/web-based learnings to game-based, podcasts and on-the-job learning.
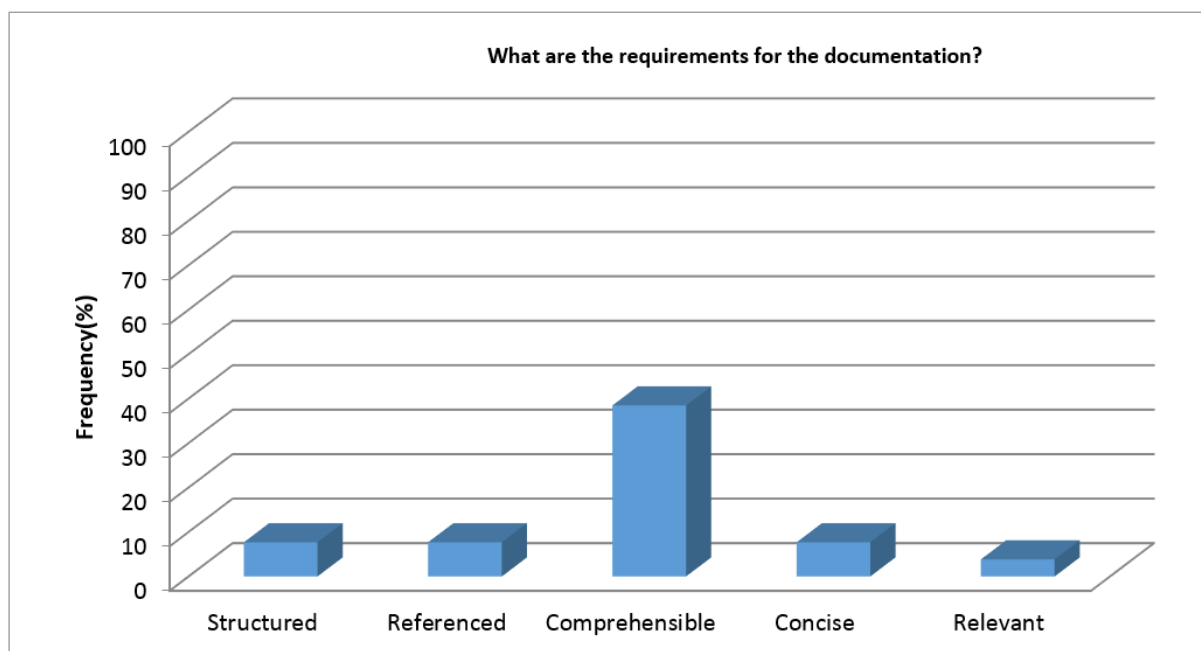


**Figure 4:** 12th question ("What are the requirements for the documentation?").
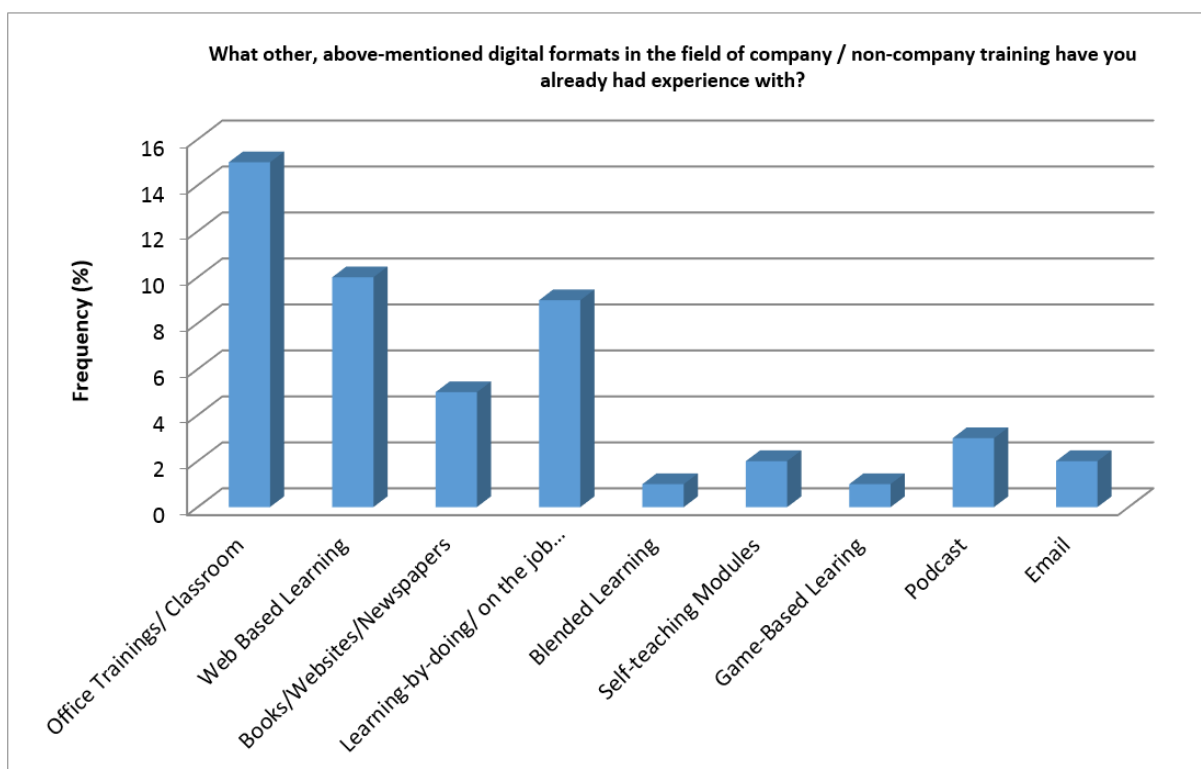


**Figure 5:** 18th question ("What other, above-mentioned digital formats in the field of company / non-company training have you already had experience with?").

The integrated analysis of responses shows that, although the questionnaire provides extensive insight into the auditor's needs and the application of requirements, there is a clear need to go further with the requirements analysis through using other methods. For that purpose, we organised

a brainstorming session across the Germany, Switzerland and Austria (GSA) Innovation Management of EY to increase the amount of feedback and to pin down the needs of junior and senior professionals in an efficient manner.

## 2.3 Brainstorming session

Based on the questionnaire analysis of responses we continued the requirement analysis with conducting a brainstorming session within the GSA Innovation Management team. The brainstorming session took place at a Germany-Switzerland-Austria internal EY level. The result of this session focuses on the specific needs of auditors and the audit process, which needs to comply with the International Standard(s) on Auditing (ISA). The main outcome of the brainstorming session is that the MOVING platform should support the ISA requirements. The most relevant standards discussed have a direct application within the scenarios presented in the DoA and are extensively detailed in Section 4 of this deliverable. The ISA 240 on text mining and search for unusual journal entry description is directly connected to the PEST Analysis in Scenario 1. The auditor assigned with the task to identify potential risks of economic and financial changes in laws and regulations as part of the SWOT analysis can apply here the technique of the tag cloud. Moreover, we see a direct link between the procedures described in ISA 550 on related party search and ISA 315 on understanding the entity and the environment, and the risk assessment performed by the auditor in Scenario 1. Furthermore, the procedures of ISA 315 can play an important part in Scenario 2 focused in innovation of advisory services. ISA 500 on scanning other information available is similar to ISA 315 and the procedures described in this standard can be important to both scenarios. The requirements of ISA 720 requires the auditor to read other information which is included in a document containing audited financial statements and the auditor's report thereon on laws and regulations changes concur with Scenario 1, focused on the compliance to European laws and regulations, while those of ISA 230, on quality of audit documentation, support both scenarios.

## 2.4 Related work

In a time of rapid technology advancement and exponential knowledge growth, there is a high probability that tools might already exist to help reaching a proposed goal or enabling learning techniques. Therefore, for the analysis of previous work, we performed a text mining and graph databases tools review. Our tools research concluded that the variety of existing tools do not meet the basic needs of the MOVING platform user. We describe in the following subsections a competitor's analysis of available tools and techniques that exist on the market or are open source, as well as a literature review on the topic.

### 2.4.1 Graph databases

One important feature resulting from the requirement analysis is the use of network graphs for visualisation and analysis of results. There are various graph databases available to use. Graph

databases are for far more than just basic visualisation. Insight and intelligence drawn from graph data connections supports success of business digitalisation and success.

In 2015, Forrester published a review discussing graph databases and graph visualisation software (Yuhanna, 2015). In the review, they put a focus on business use cases for the application of graph database technology and vendors offering graph database services. Figure 6 shows typical graph database vendors and specifications of their product and services such as licence type (commercial or open source), query language and top use case applications (Google's Cayley[2] and Ontotext's GraphDB[3], are not included in the figure).

NorthData[4] is an analysis platform that uses commercial register notices and other publications of German companies to gain economic, financial and interrelationship information. Moreover, it uses big-data processing, artificial intelligence and customised graphs for the visualisation.

| Vendor | Initial release | Open source/ commercial | Graph model | Query languages | Top use cases |
|---|---|---|---|---|---|
| Complexible | 2012 | Commercial | RDF | SPARQL | Complex analytics, 360-degree view of the customer, reference data, and data provenance apps |
| DataStax | 2011 | Open source | Property | Gremlin | Advanced analytics, fraud detection, intelligence, and social networking |
| FlockDB | 2010 | Open source | Property | Java | Social networking, fraud detection, and complex analytics |
| Franz | 2005 | Commercial Open source | RDF | SPARQL, RDFS++, OWL2-RL, Prolog | Healthcare analytics, intelligence and defense, and publishing and banking apps |
| Neo Technology | 2007 | Open source | Property | Cypher, native API, TinkerPop | Real-time recommendations, graph-based search, network and IT operations, identity and access management, fraud detection, and MDM |
| Objectivity | 2011 | Commercial | Node and edge objects | Java | Security/threat detection, fraud, complex manufacturing, infrastructure management, and predictive analytics |
| Oracle | 2015 | Commercial | Property | Java, Gremlin, Groovy, Python | Advanced analytics, social networking analytics, knowledge networks, fraud/anomaly detection, and recommendation systems |
| Orient Technologies | 2011 | Open source | Property | REST, Gremlin, SPARQL, SQL | Social networks, recommendation systems, fraud detection, and linked data apps |

121473     Source: Forrester Research, Inc. Unauthorized reproduction or distribution prohibited.

**Figure 6:** Graph databases vendors (Yuhanna, 2015).

---

[2] https://cayley.io (2017-03-27)

[3] http://graphdb.ontotext.com (2017-03-27)

[4] https://www.northdata.de (2017-03-27)

## 2.4.2 Graph visualisation

There is a variety of graph visualisation frameworks available on the market with various features and different user-friendly levels, for both "back-end" developers and "front-end" users. Neo Technology (Neo4j)[5] (Figure 6) offers the most commonly used graph database, which is a property graph type and uses the Cypher[6] query language. Among the visualisation tools the Neo Technology proposes are Tom Sawyer[7], a powerful Software Development Kit (SDK), along with others such as Linkurious[8] and KeyLines[9], both Java-based visualisation tools. Moreover, Linkurious also offers an SDK product; nevertheless, it is less user-friendly than Tom Sawyer is. Tom Sawyer is a combination of JavaScript (JS) programming and ready-to-use built-in features. A number of other tools available for graph visualisation include Gephi[10], Popoto[11], Miru[12] and Structr[13]. We reviewed the aforementioned tools for their visualisation capabilities and user-friendliness. Although they all offer a variety of features, the tools require advanced programming skills for the understanding of functionalities and advanced knowledge about graph databases. The goal of MOVING is to provide researchers and practitioners with fast and user-friendly solutions. This should help them in the application of procedures according to the two scenarios described above. Graph visualisation capabilities and further analysis based on these visualisations are part of the requirements of the ISA standards.

## 2.4.3 Big data text analytics

Text and data mining involves the use of a set of continuously evolving research techniques enabling almost anyone, with the right skills and access to data, whether text, numbers, images or in any other form, to explore that data in search of new insights and knowledge (as defined in a report conducted by the European Commission – Hargreaves et al., 2014). Text mining and analytics can help auditors improve their research results. Moreover, it helps professionals when following the procedures of the ISA standards. This applies to both scenarios described in the current use case. Figure 7 shows different types of text analytics components and capabilities, along with examples of text analytics platform features. These features are consistent with the requirements of our use case

---

[5] https://neo4j.com (2017-03-27)

[6] http://www.opencypher.org (2017-03-27)

[7] https://www.tomsawyer.com (2017-03-27)

[8] https://linkurio.us (2017-03-27)

[9] https://cambridge-intelligence.com/keylines (2017-03-27)

[10] https://gephi.org (2017-03-27)

[11] http://www.popotojs.com (2017-03-27)

[12] https://kajeka.com/miru/miru-about (2017-03-27)

[13] https://structr.org (2017-03-27)

scenarios. Auditors, when following procedures of the ISA standards throughout their work, can utilise all these.



**Figure 7:** Text analytics components and capabilities (Evelson, 2015).

A more detailed overview of vendors offering text-mining services is displayed in Tables 2 and 3, which provide insight into the software variety available for users. The two tables mainly describe the components and capabilities of the products offered by the tools/vendors listed. Some of the main problems of text mining come from the complexity of the explicit and implicit extraction concepts and of techniques such as natural language processing, feature extraction, text-based navigation, search and retrieval, categorisation and clustering (Gupta & Lehal, 2009).

### 2.4.4 Business intelligence and analytics platforms
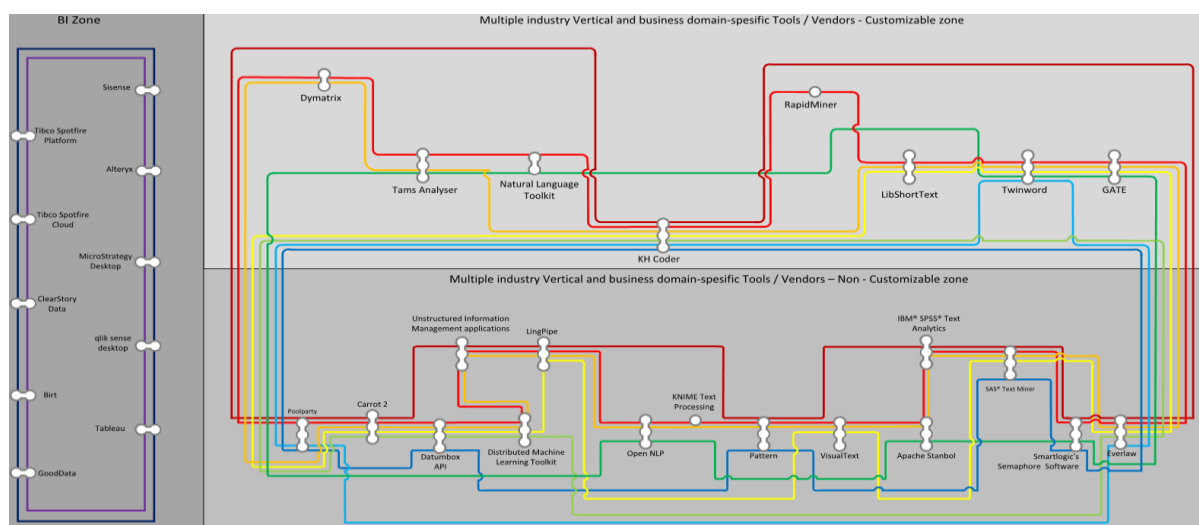


**Figure 8:** Business Analytics Chart (Evelson, 2015).

The variety of available software is overwhelming and there is a great need of a structured overview on what is out there and what type of features can be used. The chart in Figure 8 is a visual chart constructed from the review published by Forrester in 2015 (Evelson, 2015), showing tools and vendors split into three categories: BI zone, customizable zone and non-customizable zone. Moreover, Figure 9 also shows a software provider chart in the form of a quadrant plot. The plot is made of four squares where software providers are located according to technical capabilities. The provider classification contains four categories: challengers, leaders, niche players and visionaries. Based on this research users can decide on which software to invest their resources. Tables 2 and 3 provide a more detailed structure of components and capabilities, helping the user to further filter down to only features relevant to his/her research. As the tables show, the tools listed include various text mining and analytics features. However, neither of them provides an integrated platform with all available features provided. That is what the MOVING platform is pursuing: to make available an integrated platform, which a large variety of researchers and practitioners can use in their daily work.



**Figure 9:** Magic quadrant according to Gartner Review (Parenteau et al., 2016).

**Table 2:** Tools/vendors overview of text mining components and capabilities - part 1.

| Tools / Vendor | BI Zone (Tools capable of Analysing numbers) | Customizable | Non-Customizable | Multilanguage capabilities | Linguistic text preprocessing, NLP | Sub-document-level text analytics | Document-level text analytics | Cross-document text analytics | Tagging | Relationship maps | Advanced functionality | Financial Analytics | Visual Analytics |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dymatrix -DynaMine | | X | | | X | X | | | | | | | |
| Tams Analyser | | X | | | X | X | | | X | | | | |
| KH Coder | | X | | X | X | X | X | X | | X | X | | |
| Natural Language Toolkit | | X | | | X | | | | X | | | | |
| RapidMiner | | X | | | X | | | | | | | | |
| LibShortText | | X | | | X | X | X | | | | | | |
| Twinword API | | X | | | X | X | X | | X | X | | | |
| GATE | | X | | X | X | X | | | X | | | | |
| Poolparty | | | X | X | X | | | | | X | X | | |
| Carrot 2 | | | X | X | X | X | X | X | | | | | |
| Unstructured Information Management applications | | | X | X | X | X | | | | | | | |
| LingPipe | | | X | X | X | X | X | | | | | | |
| Open NLP | | | X | | X | X | | | X | | | | |
| KNIME Text Processing | | | X | | X | | | | | | | | |
| Pattern | | | X | X | X | X | | | | | X | | |
| VisualText | | | X | | X | X | X | | | | | | |
| Apache Stanbol | | | X | | X | X | | | X | | | | |

**Table 3:** Tools/vendors overview of text-mining components and capabilities - part 2.

| Tools / Vendor | BI Zone (Tools capable of Analysing numbers) | Customizable | Non-Customizable | Multilanguage capabilities | Linguistic text preprocessing, NLP | Sub-document-level text analytics | Document-level text analytics | Cross-document text analytics | Tagging | Relationship maps | Advanced functionality | Financial Analytics | Visual Analytics |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IBM® SPSS® Text Analytics | | | X | X | X | X | | | | | | | |
| Datumbox API | | | X | | X | X | X | X | | | X | | |
| Distributed Machine Learning Toolkit | | | X | | X | X | X | X | | | | | |
| SAS® Text Miner | | | X | X | X | X | X | | | | X | | |
| Smartlogic's Semaphore Software | | | X | X | X | | | | X | | X | | |
| Everlaw | | | X | X | X | X | X | X | | X | X | | |
| Ranks NL | | | X | | X | | X | X | X | | | | |
| Birst | X | | | | | | | | | | | X | X |
| ClearStory Data | X | | | | | | | | | | | X | X |
| Tibco Spotfire Cloud | X | | | | | | | | | | | X | X |
| Tibco Spotfire Platform | X | | | | | | | | | | | X | X |
| Tableau | X | | | | | | | | | | | X | X |
| qlik sense desktop | X | | | | | | | | | | | X | X |
| MicroStrategy Desktop | X | | | | | | | | | | | X | X |
| Alteryx | X | | | | | | | | | | | X | X |
| GoodData | X | | | | | | | | | | | X | X |
| Sisense | X | | | | | | | | | | | X | X |

# 3 Methodology use case 2: managing and mining research information

This section describes the methodology of the requirement analysis of the use case "Managing and mining research information", where interviews at Technische Universität Dresden (TUD) were conducted on information retrieval strategies of young researchers with regard to the scenarios described in the DoA. First in Section 3.1, the second use case as described in the DoA will be mentioned. Then in Section 3.2, the interviews for the requirement analysis and all related actions will be presented. In this section, there will be information about the interview guide, the sampling for the interviews, the methods of analysis, i.e. qualitative content analysis and the formulation of user stories as well as the primary persona we derived from these interviews. Finally, in Section 3.3 related work will be presented, i.e. two external empirical studies on the usage of digital and online-based tools (Section 3.3.1) and on requirements of researchers for the management of research data (Section 3.3.2).

## 3.1 Use case description of use case 2

The reception of the activities and results of other scientists in a certain research field as well as a precise knowledge of the current state of the discussion within the relevant scientific community is one of the core tasks in scientific work. On the one hand, this practice makes sure that newer or even older ideas cannot be falsely stolen and solutions do not have to be worked out twice. On the other hand, it is the base for the scientific principle that new scientific ideas should always build on the results of previous work to enhance scientific innovation. Therefore, serious scientific work has to go hand-in-hand with good literature review.

Today, the internet is the central place for scientific information and literature review. The available sources are various: in addition to increasing library catalogues, online archives and databases as the classical approaches, social networking sites (e.g. ResearchGate[14]) and content sharing platforms such as SlideShare[15] or scientific blogs are available as sources. Since not only the content but also the number of possible sources increase rapidly, researchers are also increasingly faced with the problem of information overload: information is plentiful, but there is not enough time to systematically locate and evaluate the vast information resources. The discrepancy between the theoretically offered wealth of information and the practical possibilities of their exploitation within the everyday academic work is steadily increasing and successful research today is demanding a comprehensive international perspective, more than ever before.

---

[14] https://www.researchgate.net (2017-03-27)

[15] https://www.slideshare.net (2017-03-27)

Therefore, the automated analysis of large literature corpora in a specific field of research is of great interest for scientist of every discipline. Specifically, the following fields of application are possible:

- State-of the-art on a research topic.
- Development of a (hitherto unknown) research field (e.g. in the context of qualification work or for the purpose of meta-research).
- Analysis of provenance and the genesis of discourses.
- Identification of research desiderata.
- Finding suitable partners for research projects that are active in the respective research fields.
- Actor analysis/network analysis (including for areas of scientific policy advice and strategy development).
- Strategic decisions in the field of science funding (ministries, university administrators and others).

Especially in the humanities and the social sciences, there is often a lack in understanding the principles and methods of data mining. The consequence of this is that on the one hand awareness of the possibilities is missing. On the other hand, the competencies to make meaningful use of the results are not sufficient. Below, we describe an example series of scenarios through the use case of managing and mining research information.

### 3.1.1 Scenario 1: state-of-the-art on a research topic

Ms. Brown is a junior researcher in the field of Science and Technology Studies. She is interested in the emerging interdisciplinary research field of the "Internet of Things" and therefore she wants to get an overview of existing research results and publications. For this purpose, she uses the MOVING platform. Ms. Brown is opening the website and is writing different search terms such as "Internet of Things", "Semantic Web", "networked things" and "Smart City" into the search slot of the MOVING platform. With each entry, the system suggests other semantically related keywords she could use additionally or instead. Before starting the search process, the platform is requesting Ms. Brown to define more precisely the resources that shall be included and the type of results. As included resource provider, she is choosing the content of Deutsche Zentrale Bibliothek für Wirtschaftswissenschaften (ZBW), GESIS and public content crawled from the web. With respect to the kind of results, she is selecting "open access publications" and "social media". Moreover, she puts a check for "headlines", "abstracts", "keywords" and "full text" to receive a wide overview. Afterwards, Ms. Brown selects "start search". With the list of results, she is dissatisfied: 50,000 results she can impossibly check by herself. She is uncertain what to do next. Database research usually ends up here. So she has no idea or experience, how to filter the relevant results out of the 50,000. In the navigation bar on the right, she is discovering a box called "What other people may have done". Navigating to this box Ms. Brown has yet to define from which discipline (social sciences, natural sciences, life sciences) and/or field of working (academia, economy, administration) the users should come, whose strategies and experiences at which she wants to have a look. This is important,

because due to different working and scientific culture each might have some different strategies and methods. She chooses "social sciences/academia". One proposal, among others, is "visualise results" using a network of topics. Ms. Brown is selecting "visualise results" and first is watching the proposed 2-minute "How-to Video", which was created by another user in order to explain the procedure of visualising the results and give examples when and for what purpose to use this functionality. Ms. Brown is recognising while using the visualisation function it is possible to get the various results displayed as a network structure. For example: which results are connected with which keyword, which results come from what scientific discipline, of what kind are the results, how old are the results and more. All of the above she can simply check by zooming in and out of the network and by an analysis of the visualised nodes within the network and its change over time. Ms. Brown is now one-step ahead. She is highlighting several areas within the network that are interesting for her and is selecting "show results". This time the results are about 200. Since this is still pretty much, she is limiting now the selection of searched resources and is narrowing the kind of results. To get an overview of current research questions in the field of "Internet of Things", Ms. Brown wants to have a closer look on the results of blogs, websites and Twitter posts. She then visualises results again with focus on the frequency certain terms in relation to the key term "Internet of Things" have and their change over the last 2 years period to make a trend analysis. This is another suggestion from "What other people might have done", but this time presented in the "How-to wiki", which is organised along typical questions like "What to do when receiving too much results?" and filled by other users describing their successful solutions there.

### 3.1.2 Scenario 2: finding suitable partners for research projects that are active in the respective research fields

After a few days that Ms. Brown used to read on to the topic and to develop a specific research-question for herself, she is realising that she can only work on that topic in an interdisciplinary team with other researchers. Again, she uses the MOVING platform to find potential project partners. First, she is selecting in the top navigation bar "Connecting with others". Afterwards the MOVING platform is proposing her to load her former query (the entire search process) and apply it to the current search for other researchers. The keywords, the chosen databases and kinds of results are the same. Ms. Brown is selecting "apply". Instead of a search list, immediately the visualisation will appear, as she has specified this last time as preferred presentation-mode of the results. Since the last query has been used, the marked areas she created within the network of results last time are still visible. Ms. Brown moves the mouse over the areas and different names pop up. Due to the mapping of the search results to the respective scientific discipline within the visualisation, she can check from which disciplines she should have researchers in her project to work on the research question. Using the mouse, she is choosing five scientists. Two are already in her social network, two have an account on the MOVING platform, and one is from outside. Because Ms. Brown is not very well versed with the platform, she looks again into the box "What others have done". There she is getting the recommendation to "contact people". Ms. Brown is watching the explanatory video about how to send messages via the MOVING platform and then is sending a MOVING message (a so-called MOVE)

to the researchers registered at the platform. The researcher who is not available through the MOVING platform is receiving an e-mail from the platform. The two researchers that are already in her network, Ms. Brown calls out of MOVING via Skype (or a similar instant messaging tool) and invites them to join the platform, too.

### 3.1.3 Scenario 3: strategic decisions for deciding to go for which research funding for my topics

After Ms. Brown now has a project consortium, she and the other researchers need an additional financial budget to work on the research question. Ms. Brown is thinking about from where to get funding. Again, she uses the MOVING platform. She is selecting "Funding" in the navigation and is realising the more often she uses the platform, the less extensive are her queries. Again, since MOVING uses the saved query-data, she has to specify only the type of funding that should be taken into account. Ms. Brown is immediately selecting "all", which means not only classical funding instruments of national funding agencies, EU programs, scholarships and so on, but also funding formats such as crowdfunding or tenders of companies. She is starting the search process via "Start search". Again, the visualisation is displayed, but this time she can choose between two formats of representation: either current funding announcements or experiences of other users with the different types of funding. Since Ms. Brown is still not well experienced with funding, she is choosing the visualisation-format "experiences of other users". There she can see to what of her keywords which funding-formats others used in advance. She determines that one of her consortium-members has experiences in "crowdfunding" in the area of "Internet of Things". Two of her consortium-members are also familiar with EU funding. Thus, she now wants to know which current EU funding announcements could fit to the project idea. Therefore, she switches to that format of presentation and learns that there are a couple of different announcements where the topic "Internet of Things" and "Crowd Funding" is addressed. However, a closer reading of the announcements reveals that there is no call with a good fit now as this topic had been in the focus of funding some months ago. She therefore comes back to her "trend analysis" on MOVING of the very beginning to understand more about the development of discourse for "Internet of Things". On this basis, she selects two new keywords and starts a new search. Now she finds an appropriate funding instrument: crowdfunding could be a good starting point.

### 3.1.4 Scenario 4: accompanying training materials, courses and tutorials

Since Ms. Brown is interested in crowdfunding, she selects in the navigation "training". The search result-visualisation remains the same, but now additional learning opportunities, e.g. for the different funding schemes are displayed. She realises that there is a MOOC with the topic "Alternative ways of research funding". There is also an explanatory MOOC on "Crowdfunding - 10 easy steps to get started" and a webinar on experiences with crowdfunding is offered. Ms. Brown decides to watch the MOOC and to join the webinar. In addition, all videos were automatically fragmented and annotated by the MOVING platform with the user-selected topics from his previous

query. Therefore, the user can either directly jump on the chapter of the webinar relevant to her specific interest, or watch all of it. Since Ms. Brown is very satisfied with the MOVING platform and she manages several Masters' theses where new scientific work practices also play a role, Ms. Brown recommends the platform to her students. When using the platform first they are offered a number of tutorials explaining the platform features and the main principles of operation. Of course, tutorials can be called later again. Moreover, the students have the chance to join the "young scholar community" where they can share experiences with applying data mining techniques for literature research and discuss what this means for their research and research results and even meet more experienced data research professionals. If they still do not feel familiar with using the platform, special self-learning materials are offered. Finally, yet importantly, the community allows building research tandems, bringing together experienced researchers with young ones.

## 3.2  Interviews

The requirements analysis focuses on the Humanities and Social Sciences, because within these scientific disciplines, searching for and working with literature plays a significant role. For this reason, technological support and training of young researchers seems necessary. For the requirements analysis nine guided interviews were conducted with young researchers from the Humanities and the Social Sciences at TU Dresden. The interviews have been carried out face-to-face or by telephone and have been recorded and transcribed.

### 3.2.1  Interview guide

To evaluate the requirements for the MOVING platform an interview guideline has been developed regarding use case 2 "Managing and mining research information" and the respective scenarios of the project. The first part of the guideline focused on questions about approaches, strategies and experiences of young scientists with unstructured data and information, such as texts, pictures or videos. As an introduction into the topic, the interviewees were asked what the increasing availability of information on the internet meant to them and their personal approach in general. Furthermore, young scientists were asked about the type of digital or online available information they were using, and they should then describe specific workflows of search for information on (a) state of the art on a specific topic, (b) search for funding possibilities, and (c) search for project partners for joint research projects. In addition, the interviewees should describe how they perform a search for information on legal conditions, conferences, scientific events, summer schools, scientific fairs, stays abroad, scientific exchange and online tools for research. Afterwards, concrete methods for the searching for and preparation of information and the usage of digital/online-based tools were asked. To conclude this part of the questionnaire, the interviewees were invited to name needs and suggestions for improvement.

The second part of the interview guideline focused on information literacy and the usage of digital/online-based learning offers. This part aimed on the evaluation of requirements for scenario four "Accompanying training materials, courses and tutorials" of the second use case. The German

framework of reference on information literacy is used as theoretical background (Klingenberg, 2016a). This framework was developed to standardise the teaching of information literacy within different institutional learning offers, to have similar learning goals and to be able to develop methods and concepts to reach these goals (Franke, 2016, p. 22). The framework on information literacy was first developed for education in schools (Franke, 2016, p. 23) and has been adapted to higher education in 2015 (Franke, 2016, p. 23). It consists of the following competence groups and respective activities and criteria:

- Searching: expressing the need for knowledge, finding and choosing resources, segregating information.
- Verifying: relevance of the topic, factual and formal accuracy, completeness.
- Knowledge: formulating, comparing, integrating, organising.
- Delineating: simplicity, semantic redundancy, cognitive structuring, cognitive conflict.
- Sharing: clarifying terms of use, marking citations, naming resources, using networks.

First, the framework of reference was explained to the interviewees. Then they were asked if and how they already took part in trainings on information literacy or one of the competence groups. Thereby, they were asked about trainings within their studies or beyond the context of higher education explicitly. Afterwards, the framework focused on different types of trainings. Therefore, the interviewees should describe their training behaviour in general and were then asked for their usage of digital training offers like blended learning, game based learning, MOOCs, self-study modules and web based trainings. Additionally, the interviewed young researchers were asked how useful digital training formats are for their individual training in general.

### 3.2.2 Interview sampling

To become a significant interview sample, those cases meaningful for answering the research question (Döring & Bortz, 2016, p. 303) were selected. Selection criteria was:

- Discipline: Young researchers from the Humanities and Social Sciences according to the sub-disciplines defined by the German Federal Statistical Office (DESTATIS, 2015) as far as these sub-disciplines are represented in the School of Humanities and Social Sciences of TU Dresden[16].
- Degree/status: Young researchers i.e. advanced Master students (ideally writing their Master thesis) and PhD students.
- Gender: It was paid attention that female and male researchers are both considered in the interview sample proportionally 1:1.
- Meaning of digitisation for the individual working environment (referring to department, institute, research place or study): On the one hand, interviewees for whom digitisation

---

[16] https://tu-dresden.de/gsw?set_language=en (2017-03-27)

seems to be less important have been included and on the other hand, young researchers for whom digitisation plays a more important role have been included into the interview sample. The aim was to reflect the differences in usage intensity and meaning of digital tools for the scientific work of young researchers. Possible criteria to find interviewees for whom digitisation is more important could be:

- Thematic or methodical intersections of departments, research places or studies to digital technologies or processes of digitisation, in a narrow sense (e.g. respective studies, research projects or an individual research focus); just a general (broader) relation to technology e.g. technical sociology or technical history was not considered as sufficient enough.
- Visibility of the importance of digitisation on the web, e.g. on their personal page on the websites of TU Dresden.
- Visibility and possibility to get in contact: only young researchers were selected who published their contact details (E-Mail-Address) on the websites of TU Dresden.

With the help of the named criteria, the interview sample was selected. Due to the lack of young researchers in some disciplines and missing contact data, some criteria could only be partially fulfilled. This was the case e.g. for the disciplines Sociology and Philosophy.

As a result, the following interview sample of nine interviewees was selected:

**Table 4:** Interview sample.

| No. | Discipline | Intersections to digitisation | PhD-students | Master-students | Gender |
|---|---|---|---|---|---|
| 1 | Pedagogics | X | X | | Male |
| 2 | History | X | X | | Male |
| 3 | | | | X | Male |
| 4 | Communication sciences | X | X | | Female |
| 5 | Political sciences | | | X | Male |
| 6 | Language, literature and cultural sciences | X | X | | Female |
| 7 | | | | X | Female |
| 8 | Economics | | | X | Female |

| No. | Discipline | Intersections to digitisation | PhD-students | Master-students | Gender |
|-----|------------|-------------------------------|--------------|-----------------|--------|
| 9 | Art and music | | X | | Female |

### 3.2.3 Qualitative data analysis

Initially a computer-assisted qualitative data analysis of the nine interviews was conducted. Because of the small number of cases, the method of qualitative data analysis according to Mayring (2010) has been used, but without conducting quantitative analysis steps. It was especially important to analyse the textual components in relation to the textual context within the specific interview. Thereby a significant advantage of the method of qualitative data analysis is the possibility to consider different communication contexts. Mayring (2010) suggests a mixed method approach, including qualitative and quantitative methods of analysis (Mayring, 2010, p. 48). Because of the small number of cases, only qualitative steps of analysis were performed.

The qualitative data analysis was conducted using computational methods. The usage of analysis software offers advantages in the management of interview texts, the coding and organisation of text segments, the visual representation of interrelations between text segments e.g. hierarchies, networks, and the detection of overlapping and argumentation patterns (Kuckartz, 2010, pp. 12). In addition, a computer-assisted analysis facilitates the attachment of memos and comments to codes and text segments (Kuckartz, 2010, pp. 12). The nine conducted interviews have been analysed with the help of the analysis software MAXQDA[17].

As a first step, the interviews were imported as text documents into MAXQDA. Then the coding unit were defined. The interview study aimed at the evaluation of implicitly and explicitly expressed needs of potential users of the MOVING platform. Therefore, the minimum coding unit was defined as sentence or phrase. Subsequently, the following main categories were defined: (a) literature research (scenario 1: state-of-the-art on a research topic), (b) searching for project partners (scenario 2: finding suitable partners for research projects that are active in the respective research fields), (c) searching for funding (scenario 3: Strategic decisions for deciding to go for which research funding for my topics), and (d) training (scenario 4: accompanying training materials, courses and tutorials). These categories were derived from the four scenarios of the second use case "Managing and mining research information".

For every main category, sub-categories were assigned, which have been derived from the framework of reference for information literacy (Klingenberg, 2016a). The three mentioned partial

---

[17] https://www.maxqda.com (2017-03-27)

competencies could also be understood as particular steps of the conduction of a searching process. Therefore, the following sub-categories were defined for each main category: (a) searching, (b) verifying, (c) generating knowledge, (d) delineating, (e) sharing.

During the coding process, the coders could define new sub-categories inductively or further differentiate existing sub-categories. If necessary, it was also possible to define new main categories as further use-case-scenarios. The differentiation of categories has been limited to the third sub-level. To be able to consider all expressed user needs precisely, after the first coding, a second coder reviewed the first results. The resulting adjusted category scheme can be found in annex 9.2.

### 3.2.4 User stories

Starting with the coded interviews as a second step, expressions of needs were identified and user stories have been formulated. The method of user stories is applied in agile software development to evaluate requirements from a user's perspective and develop alongside user's needs (Wirdemann, 2011, p. 51). Therefore, story cards are used to keep the core requirements from the user perspective in one sentence. The first formulation of requirements is thereby very broad and open. Specifying requirements too early is avoided intentionally (Wirdemann, 2011, pp. 53). Later developers and potential users interact on every user story and specify the requirements together as technical requirements. Therefore, generally the back of the story card is used (Wirdemann, 2011, pp. 51). Communication between potential users and developers is central for the method of user stories, the story card is understood as "promise" of the developers "to discuss a story in detail when it becomes concrete" (Wirdemann, 2011, p. 52). As a third, a group of users and developers defines step criteria of acceptance together. The criteria of acceptance are evidence what has to be implemented and how (Wirdemann, 2011, p. 55)

The user story method was used for the requirements analysis, because it was obvious that the expectations of researchers from the Humanities and Social Sciences will be relatively imprecise and will be related to experiences with other tools. Therefore, the focus of the analysis was one the hand the working steps mentioned above and on the other hand statements of the interviewees about where and how they needed or wanted support. The requirements presented in D1.1 are based on the first step of the user story method. Out of the coded interview, material user stories were identified or expressions of users' needs were rephrased to improve readability. Thus, overall 412 user stories could be identified from the nine interviews. Most of them are related to scenario 1 "state of the art on a research topic" covering the entire range of partial competencies of the framework of reference for information literacy. Regarding scenario 2 and scenario 3, it became obvious that the interviewed young researchers have only limited experiences in searching for funding opportunities and project partners. The user stories related to scenario 4 were further analysed in Task 2.2 (Development of curricula) and the results will be presented in D2.1 (Section 3). Then the 412 user stories have been concentrated and double responses were removed. In the end, 126 user stories remained (see Section 5). Subsequently, epics were defined for different groups of user stories. Epics are "large user stories" (see ibid. p. 60), differentiated with the help of particular

user stories. As theoretical framework for the definition of epics and super epics, the framework of reference for information literacy was used again. Thus, the user stories were assigned to 12 epics, which have been related to three super epics, as can be seen in Table 7.

**Table 5:** Overview super epics, epics and number of user stories.

| Super Epic | Epic | Number of user stories |
|---|---|---|
| Search on the MOVING platform | Express the need for further knowledge. | 13 |
| | Conduct query. | 20 |
| | Refine search results. | 24 |
| | Verify completeness of the query. | 6 |
| | Select and look at several search results. | 10 |
| | Verify factual accuracy of several search results. | 3 |
| | Document the search process. | 6 |
| Generate knowledge out of information | Formulating knowledge. | 3 |
| | Compare information with other information. | 4 |
| | Integrate information into a given context. | 4 |
| | Organise - connect several information in a meaningful way. | 14 |
| Community build-up | Sharing. | 9 |
| | Communication and collaborative work. | 10 |

### 3.2.5 Primary persona

According to Noessel et al. (2014), a primary persona is the main target of an interface design. The goal of this persona is to have all needs and goals of all users combined. For the second use case, we developed throughout the interviews a primary persona, which we present alongside the scenarios in the DoA and with relation to the described persona in the DoA.

**Table 6:** Primary persona of second use case.

| Name and Age | Ms. Sally Brown, 26 years | |
|---|---|---|
| | | 18 |
| Education | Master's Degree in science of education and technology. | |
| Professional background | Research assistant for one year in the field of Science and Technology Studies. | |
| Information retrieval behaviour on new topics in research | • Looks up key words in library catalogue while conducting a Google and Google Scholar search at the same time.<br><br>• Tries to find monographies about the topic and uses the snowball-method (i.e. she looks into the citations and references of articles and books and uses the cited authors and resources for further literature research, this strategy is similar to the "snowball sampling" a sampling method in qualitative empirical research (Döring & Bortz, 2016, p. 308))<br><br>• Looks up similar key words to her search.<br><br>• Searches in journal databases and social media (e.g. Twitter) for current debates on the topic.<br><br>• Delimits the key words she is interested in according to the current focus.<br><br>• Searches again in journal articles and reads now the abstracts.<br><br>• Saves the found references in the reference management system ZOTERO with which she is working since her master's thesis.<br><br>• Documents her search with taking notes on her laptop in a simple | |

---

18 Osorio, F. (2016). University Life 294 [Photo]. Retrieved from
https://www.flickr.com/photos/francisco_osorio/30359401243/ (2017-03-27)

| | |
|---|---|
| | note program (Evernote[19]) so that she can trace back, on which key word and topic she was searching for at which time. |
| Information retrieval behaviour on finding project partners | • Has weak experience with finding project partners.<br>• Looks up up-to-date journal articles on the specific topic and searches for frequently occurring authors.<br>• Would ask experienced colleagues about the authors she has found, whether they know them or not.<br>• Would contact them via email. |
| Information retrieval behaviour on finding funding opportunities | • Has weak experience with finding funding opportunities.<br>• Asks colleagues which organisations nationwide and on an international level fund project in the field of her interest.<br>• Looks up on the websites of these organisation which programs they have and writes them down with the respective deadlines<br>• Asks again her colleagues about the probability of being funded for a specific funding program.<br>• Decides on whether to apply or not. |
| Information retrieval behaviour on finding training materials | • Likes the idea of digital training material combined with face-to-face courses.<br>• Has made a few experiences with online courses but never finished a course.<br>• Is still relying on her courses she has attended at university for learning on how to write scientifically. |

## 3.3  Related work

In addition to the interview study, two scientific research studies on the usage of digital and online-based tools for data and information management have been analysed. In the context of the quantitative research study, "Innovations in scholarly communication", conducted under the coordination of the University of Utrecht[20], from May 2015 to February 2016 20,663 researchers from different scientific disciplines have been surveyed. The aim of the survey was to get an insight

---

[19] https://evernote.com  (2017-03-27)

[20] https://101innovations.wordpress.com (2017-03-27)

which online-based tools do researchers use and which significant changes of research processes occur[21]. Therefore, we asked questions on the usage of online-tools for the execution of 17 activities of the research process. These activities have been predefined, based on a tool analysis (for the results of the tool analysis see[22]), shown in Table 4.

**Table 7:** Different activities within the research process as basis for the survey questions on tool usage.

| Research phases | Activities |
|---|---|
| Preparation | • Project management<br>• Crowdsource, define research priorities, ideas, collaborations<br>• Fund, get contract |
| Discovery | • Search<br>• Get access<br>• Get alerts, get (reading recommendations)<br>• Reference management<br>• Read, view<br>• Annotate, tag (during, after reading) |
| Analysis | • Experiment and collect, mine, extract data<br>• Share notebooks, protocols, workflows<br>• Analyses |
| Writing | • Visualize<br>• Write text and/or code<br>• Cite<br>• Translate |

[21] https://101innovations.wordpress.com/survey-2015-2016 (2017-03-27)
[22] http://bit.ly/innoscholcomm-list (2017-03-27)

| Research phases | Activities |
|---|---|
| Publication | • Archive/share: code, data / video, publications, posters, presentations<br><br>• Present research findings<br><br>• Peer review and comment/recommend<br><br>• Select journal to submit<br><br>• Publish |
| Outreach | • Outreach/valorisation<br><br>• Researcher profiling (and social network) |
| Assessment | • Comment<br><br>• Peer review (post-pub)<br><br>• Measure impact<br><br>• Assessment (of researcher/ researcher group) |

From all predefined activities, only some have been selected for the survey, "(…) for their overall importance (…) and for their spread across the research workflow covering discovery, analysis and writing as well as publication, outreach and assessment" (Kramer & Bosman 2016, p. 2). For every activity, seven tools were predefined as possible responses as well as the possibility to name "(also) other used tools)". Additionally, questions on Open Access and Open Science were asked and finally questions on demographic data have been evaluated (Kramer & Bosman 2016, p. 2). The sampling was carried out completely open and without probability sampling. The questionnaire targeted researchers and people supporting research. Therefore, the survey was distributed on the internet via Twitter, mailing lists, blogposts, the survey website and other channels (Kramer & Bosman 2016, S. 3). Finally, 20,663 filled and valid questionnaires were used for statistical data analysis. In the context of the requirements analysis for the MOVING platform, this research study was used to get a first overview over processes of information research and possibly usable online tools.

In addition to the quantitative study above, the qualitative interview study "bwFDM"-communities (Tristram et al., 2016) is considered for the requirements analysis. Compared to the study "Innovations in scholarly communications"[23], this research study is more restricted in its coverage and thematic focus. At the centre of this study is the evaluation of requirements of researchers from universities of Baden-Wuerttemberg for the management of research data. It was especially

---

[23] https://101innovations.wordpress.com/ (2017-03-27)

important to find out, which steps have to be taken to facilitate the handling of research data (Tristram et al., 2016, p. 2). From January 2014 to June 2015, all-in-all 627 semi-structured face-to-face interviews have been conducted (overall 779 interviewed persons). The interview study took place at the level of scientific communities at universities of Baden-Wuerttemberg: "A scientific community is thereby defined as a group of researchers that work together at the same location and on closely related scientific topics" (Tristram et al., 2016, p. 4). In each case, one person was interviewed (Tristram et al., 2016, p. 4). Postdocs as well as PhD-students, IT professionals and partly BA/MA students were interviewed to get the full picture over all hierarchical levels. The interview guideline included introductory questions and questions on the topics environment of the data, data processing, data storing, publication and data sharing (Tristram et al., 2016, p. 5). Open as well as semi-closed and closed questions are used for the interview guideline. The interview results have been harmonised with the help of User Stories, which have been thematically clustered (Tristram et al., 2016, p. 5). In the context of the requirements analysis for the MOVING platform, this study is used to identify an appropriate method to prepare the results of the requirements analysis for further technical development in the project. Finally, significant knowledge can be gained regarding the general framework of the handling of information and data.

### 3.3.1 Innovation in Scholarly Communication - Study

In general, mostly persons from the Social Sciences and Economics attended the "Innovations in scholarly communication" survey and most of the participants came out of the context of higher education (Kramer & Bosman 2016, p. 4). The following presentation of the results of the study includes three of the predefined phases of the above-mentioned research process: discovery, writing and publication. A closer look is taken on the activities search, get access, get recommendations, read, reference management and archive/share publications.

The survey question on the usage of online tools for searching scientific literature, data etc. was answered by almost all of the participants of the survey (n=20,427)[24]. The following digital/online-based tools have been predefined: Google Scholar[25], web of Science[26], Scopus[27], Mendeley[28], World Cat[29], PubMed[30] and Paperity[31] (see questionnaire[32]). Additionally, it was possible to name other and/or additional used tools. Google Scholar has been named as used tool from most of the

---

[24] http://dashboard101innovations.silk.co/page/Search (2017-03-27)

[25] https://scholar.google.com (2017-03-27)

[26] http://wokinfo.com (2017-03-27)

[27] https://www.scopus.com (2017-03-27)

[28] https://www.mendeley.com (2017-03-27)

[29] https://www.worldcat.org (2017-03-27)

[30] https://www.ncbi.nlm.nih.gov/pubmed (2017-03-27)

[31] http://paperity.org/ (2017-03-27)

[32] https://101innovations.files.wordpress.com/2016/02/101-innovations-survey-english.pdf (2017-03-27)

participants (around 18,000), followed by Web of science (around 8,000) and PubMed (around 8,000 participants). In the Arts and Humanities and Social Sciences and Economics Google Scholar is the mainly used online-based tool for scientific literature research, too (2,913 and 6,057 participants). However, in the Arts and Humanities Google Scholar is followed by Worldcat as search-tool (1,034 participants), followed by Web of Science (902 participants), while Social Scientists (including Economics) prefer Web of science on the second place (2,726 participants) before Scopus (1,892 participants)[33].

Regarding the question, how the get access to scientific literature, data, etc., the participants could choose between possible answers like "access through my institution or library", "pay per view on publisher platform", ResearchGate, Research4Life, Open Access Button, Deepdyve or "E-Mail the author", further or alternative tools could also be named[34]. This question has also been answered by almost all of the participants of the survey (n=20,188). Over all disciplines, but also when focussing on the Arts and Humanities and Social Sciences including Economics, it becomes apparent, that mostly institutional access e.g. via universities or libraries is used (around 18,000 participants). In addition, ResearchGate as scientific network plays an important role to get access to scientific literature and data (around 10,000 participants). Additionally, the survey participants use the possibility to contact the author via E-Mail or use the Open Access-Button (6,000 participants each)[35].

Scientific literature and data can also be found with the help of recommendations. The respective question, how the survey participants get recommendations on scientific literature, data, etc. was answered by n=16,157 of the survey participants. They could choose among Google Scholar, JournalTOCs[36], Browzine[37], Mendeley, F1000 prime[38], Sparrho[39] and ResearchGate and name other (also) used tools (see questionnaire[40]). Again, Google Scholar is the mainly used tool for recommendations (around 10,000 participants); followed by ResearchGate (around 8,000 participants) and the category "other tools" (around 4,000 participants). In the category "other tools" participants named PubMed (554 participants), Academia.edu[41] and Twitter[42] (274 participants) as most important tools. The meaning of RSS-Feeds to get information on relevant scientific literature seems to decrease significantly; only 85 participants use this tool to get recommendations. In

---

[33]  https://101innovations.files.wordpress.com/2016/10/analysis-template-survey-data-cleaned-global-ili4.xlsx (2017-03-27)

[34] https://101innovations.files.wordpress.com/2016/02/101-innovations-survey-english.pdf (2017-03-27)

[35] http://dashboard101innovations.silk.co/page/Access (2017-03-27)

[36] http://www.journaltocs.hw.ac.uk (2017-03-27)

[37] http://browzine.com (2017-03-27)

[38] http://f1000.com/prime (2017-03-27)

[39] https://www.sparrho.com (2017-03-27)

[40] https://101innovations.files.wordpress.com/2016/02/101-innovations-survey-english.pdf (2017-03-27)

[41] https://www.academia.edu (2017-03-27)

[42] https://twitter.com (2017-03-27)

addition, Facebook[43] does not play such an important role for getting recommendations on scientific information, considering all disciplines (73 participants)[44]. The usage of online-based tools for recommendations is similarly distributed in the Arts and Humanities and Social Sciences including Economics. Google Scholar is also the most used tool in the Arts and Humanities as well as Social Sciences including Economics (1,657 and 3,478 participants) followed by ResearchGate (756 and 2,512 participants) and "other tools"[45]. As additionally or alternatively used tools Acedemia.edu was named most often (259 participants of Arts and Humanities and 275 participants of Social Sciences including Economics), followed by Twitter (58 participants in Arts and Humanities and 80 participants in Social Sciences including Economics) and Facebook (28 participants in Arts and Humanities and 30 participants in Social Sciences including Economics)[46].

With regards to how and where they read scientific literature and annotate them, the survey participants could choose between the following tools: Adobe Acrobat Reader[47], view HTML-version online, iAnnotate[48], Readcube[49], UtopiaDocs[50], Mendeley, Hypothes.is[51], etc. (see questionnaire[52]). This question has been answered by n=19,991 of the participants. Generally, the survey participants use the Adobe Acrobat Reader to read and annotate scientific literature (around 18,000 participants) or view texts as HTML-version (around 8,000 participants). Often the reference manager Mendeley (around 3,000 participants) and "other tools" (around 3,000 participants) are used. Within the category "other tools", the Apple-application Preview is the mainly used tool (438 participants) followed by Papers[53] (290 participants). It should be noted that the participants named in the mid-range online-based reference managers like Zotero[54] and Endnote[55] to read and annotate as well as the offline hard copy and Microsoft Word, within the category "other tools". Participants from the Arts and Humanities actually prefer Microsoft Word[56] (65 participants) to Preview (64 participants) and Zotero (60 participants) as "other tool"[57]. In the Social Sciences including Economics within this

---

[43] https://facebook.com (2017-03-27)

[44] http://dashboard101innovations.silk.co/page/Alerts (2017-03-27)

[45] https://101innovations.files.wordpress.com/2016/10/analysis-template-survey-data-cleaned-global-ili4.xlsx (2017-03-27)

[46] http://dashboard101innovations.silk.co/explore (2017-03-27)

[47] https://acrobat.adobe.com/us/en/acrobat/pdf-reader.html (2017-03-27)

[48] https://www.iannotate.com (2017-03-27)

[49] https://www.readcube.com (2017-03-27)

[50] http://utopiadocs.com (2017-03-27)

[51] https://hypothes.is (2017-03-27)

[52] https://101innovations.files.wordpress.com/2016/02/101-innovations-survey-english.pdf (2017-03-27)

[53] http://papersapp.com (2017-03-27)

[54] https://www.zotero.org (2017-03-27)

[55] http://endnote.com (2017-03-27)

[56] https://products.office.com/en/word (2017-03-27)

[57] http://dashboard101innovations.silk.co/page/Read (2017-03-27)

category mainly Preview is used (126 participants), followed by Zotero (97 participants). In addition, Microsoft word is relevant (93 participants) as well as the hard copy (92 participants)[58]. Out of the pre-given tools, Adobe Acrobat Reader and the HTML-view play an important role in the Arts and Humanities as well as Social Sciences including Economics[59].

Within the research phase of writing especially reference management is important. Therefore, the survey participants could choose between EndNote, Mendeley, Zotero, RefWorks[60], Papers, Citavi[61] and RefME[62] and name other additional or alternative tools (see questionnaire[63]). The question of how they manage references and which tools they use, answered 15,956 of the participants. The top-three tools for reference management are Endnote (around 8,000 participants), Mendeley (around 4,000 participants) and Zotero (around 3,000 participants). As additionally or alternatively used tools for reference management BibTex was named most often, (527 participants)[64]. Also within the Arts and Humanities as well as the Social Sciences including Economics, EndNote is the mainly used reference manager (914 participants and 2,161 participants). Social Scientists including Economics prefer Mendeley (1,240 participants) to Zotero (1,092). It is the other way around in Arts and Humanities (Mendeley: 507 participants, Zotero: 710 participants)[65]. As "other tools", participants from the Arts and Humanities as well as participants from the Social Sciences including Economics named Microsoft Word as their mainly used tool for reference management (53 participants from the Arts and Humanities, 97 participants from the Social Sciences including Economics)[66]. Among others, BibTex, Evernote and Excel[67] were named, but only with minor importance[68].

Regarding the activity of sharing during the research phase of publication, the survey participants could choose between arXiv.org[69], PubMed Central[70], bioRxiv[71], ResearchGate, SSRN[72], the usage of institutional repositories and the sharing of working papers. The participants could also name "other

---

[58] http://dashboard101innovations.silk.co/explore (2017-03-27)

[59] https://101innovations.files.wordpress.com/2016/10/analysis-template-survey-data-cleaned-global-ili4.xlsx (2017-03-27)

[60] https://www.refworks.com (2017-03-27)

[61] https://www.citavi.com (2017-03-27)

[62] http://www.citethisforme.com (2017-03-27)

[63] https://101innovations.files.wordpress.com/2016/02/101-innovations-survey-english.pdf (2017-03-27)

[64] http://dashboard101innovations.silk.co/page/Reference-management (2017-03-27)

[65] https://101innovations.files.wordpress.com/2016/10/analysis-template-survey-data-cleaned-global-ili4.xlsx (2017-03-27)

[66] http://dashboard101innovations.silk.co/explore (2017-03-27)

[67] https://products.office.com/en/excel (2017-03-27)

[68] http://dashboard101innovations.silk.co/explore (2017-03-27)

[69] https://arxiv.org (2017-03-27)

[70] https://www.ncbi.nlm.nih.gov/pmc (2017-03-27)

[71] http://biorxiv.org (2017-03-27)

[72] https://www.ssrn.com (2017-03-27)

(also) used tools" (see questionnaire[73]). Around 9,000 participants use the scientific network ResearchGate to share publications, around 6,000 participants upload their publications into institutional repositories and 3,000 participants named "other (also) used tools"[74]. In this category Academia.edu is the most common used tool for sharing (around 1,000 participants), followed by publishing on a personal website (304 participants) and sharing via Dropbox[75] [76]. Participants from the Arts and Humanities and the Social Sciences including Economics use mainly ResearchGate and institutional repositories for sharing, too[77]. Whereupon participants from the Arts and Humanities use institutional repositories more often than ResearchGate (1,103 participants named institutional repositories, but only 863 named ResearchGate). Within the Social Sciences, including the Economics, it is the other way around (ResearchGate: 2,873 participants, institutional repositories: 2,124 participants). On the third place, "other (also) used tools" are named in the Arts and Humanities as well as the Social Sciences including Economics, too[78]. Thereby, Academia.edu is also extensively used, followed by a personal web page and sharing via Dropbox[79].

### 3.3.2 "bwFDM"-Communities-Study

The results of the "bwFDM"-Communities-Study[80] are related to the subjects: structural framework conditions (copyright and data security, range of information, scientific data culture), digital processing of research data (support in close contact to research, scientific Software tools, formats and unique standards), data collection and exchange (data of non-university institutions, expectations regarding digitisation, scientific collaboration), preservation of information (documentation of projects and data, repositories for the publication of research data, archiving) and other aspects such as IT-infrastructure and support, licenses, financing, Open Science and reservations about research data management. In the following, only those results that are relevant for the requirements analysis of the MOVING platform are presented.

Out of overall 2,554 user stories, 117 could be identified, which deal with questions and problems related to copyright and data security (Tristram et al., 2016, p. 9). Both topics are relevant within the whole research process. Researchers need support, because questions regarding copyright and data security are often not part of their disciplinary education. Copyright is mainly relevant in the phase of

---

[73] https://101innovations.files.wordpress.com/2016/02/101-innovations-survey-english.pdf (2017-03-27)

[74] http://dashboard101innovations.silk.co/page/Archive-share-publications (2017-03-27)

[75] https://www.dropbox.com (2017-03-27)

[76] http://dashboard101innovations.silk.co/page/Archive-share-publications (2017-03-27)

[77] https://101innovations.files.wordpress.com/2016/10/analysis-template-survey-data-cleaned-global-ili4.xlsx (2017-03-27)

[78] https://101innovations.files.wordpress.com/2016/10/analysis-template-survey-data-cleaned-global-ili4.xlsx (2017-03-27)

[79] http://dashboard101innovations.silk.co/explore (2017-03-27)

[80] The study is part of the project "bwFDM" on research data management investigation in Baden-Wuerttemberg, for further information see https://bwfdm.scc.kit.edu/english/index.php (2017-03-27)

publication. The "bwFDM"-Communities-Study shows, that the topic copyright and data security is mainly important within the Social Sciences. 61 user stories (out of 117) could be identified from the Humanities and Social Sciences, while only 28 could be identified from Life Sciences, 22 from Engineering Sciences and only 6 from the Sciences (Tristram et al., 2016, p. 8). The authors trace this to the high relevance of personal data in the Humanities and Social Sciences. The following legal aspects could be identified, which have to be clarified: clarification of legal questions, legal regulations as research barrier and data secure infrastructure, protection and commercialisation rights for produced data. Especially the need for information and guidance in the case of legal questions is significantly high (48 out of 117 user stories) (Tristram et al., 2016, p. 9). Additionally, the need for an IT-infrastructure for the storing, sharing and archiving of research data could be identified. Explicitly named by interviewees from the Humanities and Social Sciences has been a need for data secure infrastructure for processing individual-related research data (Tristram et al., 2016, p. 10).

Regarding the need for information on the management of research data, the study shows, that there is an increased demand for information in all disciplines (Tristram et al., 2016, p. 18). Different possible information channels are mentioned without any clear preferences, from offline face-to-face trainings to online offers like web-portals or online courses (Tristram et al., 2016, p. 18). Of the total user stories amount, 121 mention the need for an online-based information portal on data management. One third of the 121 user stories mentioned especially a need for information on "definitions and methods of research data management" (Tristram et al., 2016, p. 20). Sporadically interviewees mentioned that it would be good to get software recommendations or information on database build up. Another larger part of user stories is related to information on local or regional data services (Tristram et al., 2016, p. 20). Regarding courses (84 user stories) a demand for general trainings on research data management, introductory courses etc. for young researchers as well as trainings on specific software for research data management could be identified (Tristram et al., 2016, p. 20). The need for a stronger integration of the topic of research data management into the formal curriculum of students and into the training of PhD students mentioned 15 user stories (Tristram et al., 2016, p. 21).

Of the 2554 user stories, 271 deal with the processing of research data and related needs for support. Due to the advancing digitisation and the growing importance of research data management, it is necessary to adopt IT-related competencies. A slightly increased need can be identified within the Humanities and Social Sciences, because of the absence of IT-related topics within the curriculum (Tristram et al., 2016, p. 37). In 187 user stories the topic project and data documentation has been mentioned (Tristram et al., 2016, p. 74), especially a demand for information in research data management in research projects could be identified (Tristram et al., 2016, p. 79). In addition, the interviewees mentioned a need for trainings, e.g. workshops or online tutorials, regarding that topic (Tristram et al., 2016, p. 79). The subject "Repositories for the publication of research data" was mentioned in 243 of the 2554 user stories. Thereby the exchange

of research data, e.g. as supplementary material to publications, the possibility to download research data or structured repositories was mentioned as important (Tristram et al., 2016, p. 87).

# 4 Use case requirements use case 1: research on business information by public administrators

The following subsections describe the different user stories of use case 1 that are aligned to the ISAs and are integrated in the scenarios of use case 1 (as outlined in Section 2.1). The persona of use case 1 is the professional accountant, more specifically the auditor of the financial statements who is required to comply with the ISA when using the MOVING platform. The requirements for the MOVING platform derived from these user stories are tagged with a unique ID and summarised together with the requirements derived from use case 2 (Section 5) in Section 6.

## 4.1 ISA 315 – Understand the entity and the environment

*Background: Objective of ISA 315*

ISA 315 requires the auditor to identify and assess the risks of material misstatement, whether due to fraud or error, through understanding the entity and its environment, thereby providing a basis for designing and implementing responses to the assessed risks of material misstatement (ISA 315 para. 3).

To enable the understanding of transaction classes, account balances and disclosures expected in the financial statements, the auditor should obtain an understanding of the following (ISA 315 para. 11):

1. Relevant industry, regulatory, and other external factors including the applicable financial reporting framework.
2. The nature of the entity, including: (a) its operations, (b) its ownership and governance structures, (c) the types of investments that the entity is making and plans to make, including, investments in special-purpose entities and (d) the way that the entity is structured and how it is financed.

*Application: Practical issues*

The risk assessment procedures of ISA 315 shall include (ISA 315 para. 6):

1. Inquiries of management and of others within the entity who in the auditor's judgment may have information that is likely to assist in identifying risks of material misstatement due to fraud or error.
2. Analytical procedures.
3. Observation and inspection.

Currently, in an audit of financial statements – and when performing the abovementioned procedures - most of the underlying data is quantitative in nature. The auditor addresses qualitative aspects of the entity and the environment by the inquiries held with management, however, these inquiries include the inherent risk of bias and incomplete information.

To address these risks, the auditor validates the information obtained from management with evidence from other sources of information, e.g., the entity's website and contracts. However, in today's era of big data, the qualitative information available quickly become overwhelming. A tool to enable the efficient structuring of large amount of qualitative information available to identify relevant topics, concepts, entities, locations, and persons can therefore increase the overall efficiency of the audit.

Another practical issue when performing procedures to understand the entity and the environment is that professional judgment is used when defining the nature and extent of the analysis. There are no indicators and guidance on when the analysis is finished and when the information obtained is enough for a sufficient understanding of the business.

### Application of the MOVING platform

The MOVING platform can support the auditor in understanding the entity and the environment by allowing for very detailed search settings, structuring the large amount of qualitative information available and identifying and visualising relevant concepts, entities, locations, persons as well as their dependencies within the given dataset. Due to the current practical issues regarding the visualisation of large amounts of qualitative information, specific consideration will need to be devoted to the efficient and effective visual encoding of information. When doing this, we will also address the mapping of logical to visual properties in a way that goes beyond the current state of the art and thus makes a difference to the visualisations, currently available.

The adaptive training support of the platform can support the user in determining when the analysis is completed, i.e., when the auditor has obtained a sufficient understanding of the entity and the environment, and when he/she obtained enough relevant information.

Features of the application that supporting the understanding of the entity and the environment according to ISA 315 include:

1. A faceted search allowing the auditor to search in a wide amount of data sources available and specify detailed search settings.
2. A search profile management with pre-defined search settings for different tasks and the option to save the search settings applied, import and export them.
3. Flexible visualisations that effectively encode the quantitative information available visually and make a difference to the currently existing solutions.

The detailed requirements are as follows:

**Table 8:** ISA 315 requirements.

| ID | Topic | Keyword | Requirement |
|---|---|---|---|
| **#EY001** | Faceted search | Search term | Define one or multiple search terms. |

| #EY002 | Faceted search | Search term | Exclude one or multiple search terms. |
|---|---|---|---|
| #EY003 | Faceted search | Data sources | Include or exclude certain data sources. |
| #EY004 | Faceted search | Data sources | Data sources should include both the world wide web and the possibility to define one or multiple specific websites to include. |
| #EY005 | Faceted search | Data sources | Browse for files stored on the local device to include in the analysis. Those files can be of various file formats, e.g., .pdf, .doc, .doc, .rtf, .txt, .xls, .xlsx, .csv, .htm, .html. |
| #EY006 | Faceted search | Data sources | The upload of data sources stored on the local device is restricted due to data privacy issues. |
| #EY007 | Faceted search | Data sources | Connection to various literature databases available to include publications of several disciplines. |
| #EY008 | Faceted search | Search extent | Specify the depth of the search, when applicable (e.g., title, abstracts, full text, and metadata). |
| #EY009 | Faceted search | Geographic region | Limit the search to certain geographical areas (e.g., the entity's headquarters country). |
| #EY010 | Faceted search | Disciplines | Limit the search to certain disciplines, when applicable (e.g., when selecting a database that covers multiple disciplines). |
| #EY011 | Faceted search | Dynamic interface | Based on the search settings, the remaining search criteria become enables or disabled (e.g., when a file on the local device is included that has no metadata, the depth of the search cannot be set to metadata). |
| #EY012 | Faceted search | Language | Limit the search to specific languages. By default, the language is in correspondence to the search term defined. |
| #EY013 | Search profile management | Save settings | Save search settings into search profiles or favourites for later use (e.g., the next year's understanding of the entity and the environment). |
| #EY014 | Search profile | Import/export | Option to import and export search settings in order to share them with colleagues or use them for |

| | management | settings | documentation purposes. |
|---|---|---|---|
| **#EY015** | Search profile management | Pre-defined settings | Provide pre-defined search settings for the different uses of the platform. For example, for the ISA 315 scenario, the profile should always include the company website and a hint from the adaptive training support also to include the latest management report and notes to the financial statements available (as files from the local device). |
| **#EY016** | Entity identification | (Sub-) document analysis | Entity identification in preparation for the visualisations: Extraction of entities, locations, persons and other top concepts from the data. |
| **#EY017** | Network visualisation | General | Concept graph to visualise topics around a specific keyword. The graph consists of notes and edges linking those nodes. The nodes are the most relevant concepts within the data and the edges link the concepts (nodes) that frequently co-occur within the data. |
| **#EY018** | Network visualisation | Navigation | The user should be able to navigate through the network, diving deeper into areas of interest and expanding the network in relevant directions. To increase the efficiency of the analysis, there should be an indication for the user how many documents/entities/locations/concepts will become visible when expanding the network accordingly. |
| **#EY019** | Network visualisation | Completeness | There should be an indicator of the degree of completeness on the current view. This should be based on the additional information that can be obtained by extending the network. |
| **#EY020** | Network visualisation | Filter | There should be a general option to change the displayed amount of nodes and edges. With the minimum settings, the graph only shows the most relevant nodes/edges. |
| **#EY021** | Network visualisation | Size of nodes, thickness of edges | By default, the size of the node depends on the concept's relevance (e.g., frequency of occurrence within the data) and the thickness of the edges |

| | | | depends on the number of co- occurrences of the connected nodes. We should consider adding alternatives, e.g., sizing the nodes by in-degree or out-degree. |
|---|---|---|---|
| **#EY022** | Network visualisation | Co-occurrence of concepts | Ability to specify the co-occurrence of concepts that is displayed in the graph (e.g., concepts that are connected by edges are included in the same text, paragraph or sentence). |
| **#EY023** | Network visualisation | Hovering | When hovering the mouse over nodes and edges, there should be additional information about the attributes of the node/edge (e.g., number of occurrences, data source with the most occurrences). |
| **#EY024** | Network visualisation | Statistics and measures | The network development should go hand in hand with statistical measures, e.g. centrality, concentration, density, shortest path, community clustering. These measures will be determined in more detail in D1.3: Initial evaluation, updated requirements and specifications. |
| **#EY025** | Tag cloud visualisation | General | The tag cloud displays the top keywords extracted from the data. |
| **#EY026** | Tag cloud visualisation | Size | The initial size of the tag depends on the tag's frequency within the search results. |
| **#EY027** | Tag cloud visualisation | Position | The position of the tag within the tag cloud depends on the relevance. The initial position of the tag within the tag cloud depends on the keyword's relevance. However, there should also be an option to order the tags alphabetically. |
| **#EY028** | Tag cloud visualisation | Rating of tags | When reviewing the initial tag cloud, the user can "rate" the different tags according to his research question/area of interest. Rating tags leads to a refresh of the tag cloud taking into account these ratings. |
| **#EY029** | Tag cloud | Sensitivity | There should be an option (e.g., a scroll bar) to adjust the sensitivity of the analysis. When changing |

| | | | |
|---|---|---|---|
| | visualisation | | the sensitivity, the size of the tag cloud changes as the required frequency/relevance of a concept to be included in the tag cloud changes. |
| #EY030 | Visualisations - general | Document pane | All visualisations should provide a document pane listing the documents according to the current search settings. The document pane should contain the (default) option to include all documents in the visualisation. The user should be able to select one, multiple or all documents. Upon selecting documents, all visualisations should update accordingly. |
| #EY031 | Visualisations - general | Navigation | The user can navigate through the visualisation holding the left mouse button to move through the data and using the mouse wheel to zoom in and out. Alternatively, there should be scroll bars and buttons for zooming. |
| #EY032 | Visualisations - general | Help | Each visualisation should contain a help icon where the user can obtain instructions about the current visualisation and tips how to adjust and interpret it. We can also consider integrating this feature into the adaptive training support. |
| #EY033 | Adaptive training support | Content coverage | Notify the user about the percentage of completion of the analysis for the respective search term(s) and determine when the analysis is completed, i.e. the analysis of relevant content. |

## 4.2 ISA 550 – Related parties search

*Background: Objective of ISA 550*

Many related party transactions are in the normal course of business. In such circumstances, they may carry no higher risk of material misstatement of the financial statements than similar transactions with unrelated parties. However, the nature of related party relationships and transactions may give, in some circumstances, rise to higher risks of material misstatement of the financial statements than transactions with unrelated parties. For example, related party transactions may not be conducted under normal market terms and conditions. In addition, related party transactions may operate through an extensive and complex range of relationships and

structures, with a corresponding increase in the complexity of related party transactions (ISA 550 para. 1).

The objectives of the auditor are:

1. To obtain an understanding of related party relationships and transactions sufficient to be able:

   a. to recognise fraud risk factors, if any, arising from related party relationships and transactions that are relevant to the identification and assessment of the risks of material misstatement due to fraud;

   b. to conclude, based on the audit evidence obtained, whether the financial statements, insofar as they are affected by those relationships and transactions: a. Achieve fair presentation (for fair presentation frameworks); or b. Are not misleading (for compliance frameworks);

2. To obtain sufficient appropriate audit evidence about whether related party relationships and transactions were appropriately identified, accounted for and disclosed in the financial statements in accordance with the framework (ISA 550 para. 9).

The auditor performs those procedures regarding related parties as part of the risk assessment procedures and related activities that ISA 315 and ISA 240 require the auditor to perform during the audit (ISA 550 para. 11).

***Application: Practical issues***

Multiple audit procedures address the risk arising from related party relationships and transactions:

- The engagement team discussion required by ISA 315 and ISA 240 includes specific consideration of the susceptibility of the financial statements to material misstatement due to fraud or error that could result from the entity's related party relationships and transactions.

- The auditor inquires management and others within the entity regarding:

   o the identity of the entity's related parties (including changes from prior periods);

   o the nature of the relationships to those related parties;

   o whether the entity entered into any transactions with these related parties during the period;

   o the type and purpose of the transactions and the controls that management has established;

- During the audit, the auditor maintains alertness, when inspecting records or documents, for arrangements or other information that may indicate the existence of related party relationships or transactions that management has not previously identified or disclosed to the auditor.

However, those procedures do not cover the risk of incomplete information.

***Application of the MOVING platform***

In addition to the requirements described in the previous section, the following features are relevant to the ISA 550 scenario:

**Table 9:** ISA 550 requirements.

| ID | Topic | Keyword | Requirement |
|---|---|---|---|
| **#EY034** | Faceted search | Auto-completion | A recommendation/auto-completion feature supports the specification of the search term(s). This will support the auditor in (a) searching for the correct entity and (b) identifying similar (and therefore possibly related) entities right from the beginning of the analysis. |
| **#EY035** | Faceted search | Data sources | Connection to various social media platforms to assess the entity's reputation, awareness and current news as well as potential relations between the persons and organisations identified. |
| **#EY036** | Faceted search | Period | Specify the relevant date range (e.g., the period after the entity's formation). |
| **#EY037** | Faceted search | Concept | Ability to limit the search to certain types of concepts, for example, persons and organisations only. |
| **#EY038** | Visualisations - general | Filter | Ability to limit the view to certain types of concepts, for example, persons and organisations only. |
| **#EY039** | Network visualisation | Colour | The colour of the nodes corresponds to the "entity identification" algorithm, e.g., person = red, organisation = green, location = blue, other concept = black. |
| **#EY040** | Tag cloud visualisation | Colour | The colour of the tag corresponds to the "entity identification" algorithm, e.g., person = red, organisation = green, location = blue, other concept = black. |
| **#EY041** | Tag cloud visualisation | Learning | The platform should learn from the user behaviour. For example, when a number of users rated a tag |

| | | | (e.g., a related organisation) as "high", the standard rating/weight for this tag in future search queries should adjust automatically. |
| --- | --- | --- | --- |

## 4.3 ISA 720 – Analyse other information produced by the entity

*Background: Objective of ISA 720*

Most of the auditor's work in forming the auditor's opinion consists of obtaining and evaluating audit evidence. Audit evidence is necessary to support the auditor's opinion and report. It is cumulative in nature and is primarily obtained from audit procedures performed during the course of the audit. ISA 720 requires the auditor to read other information (financial and non-financial information other than the financial statements and the auditor's report thereon) which is included, either by law, regulation or custom, in a document containing audited financial statements and the auditor's report thereon (ISA 720 para. 1, ISA 720 para. 5).

*Application: Practical issues*

Currently, the auditor addresses the requirements of ISA 720 by reading financial information, i.e., interim or annual business reports, in advance knowledge of the financial statements of the reporting period. Those annual reports usually include several hundreds of pages containing both quantitative and qualitative information in different file formats. In addition, many of that information may be sugar coated or not relevant for the auditor.

*Application of the MOVING platform*

An approach to increase efficiency in identifying and evaluating the numerous information contained in business reports can increase the overall efficiency of the audit. The requirements for both ISA 315 and ISA 550 described in the previous sections also apply for ISA 720. In addition, the following aspects should be highlighted:

- **Faceted search** data sources: In case of ISA 720, the information to analyse (business reports) are most commonly stored online on a website or on the local device as MS Word or PDF file. The MOVING platform should be able to connect to both data sources (included in requirements of ISA 315).
- **Visualisation over time** based on the document dates: The user could decide to include multiple business reports in to analysis, for example, the report of the prior period and the report from the period under audit. Thus, based on the document dates, the user should be able to compare both versions by adjusting the timeline of the network graph and the tag cloud.

**Table 10:** ISA 720 requirements

| ID | Topic | Keyword | Requirement |
|---|---|---|---|
| **#EY042** | Visualisations - general | Development over time | The network and tag cloud visualisation also show the evolution, development and degeneration of concepts over time (taking into account the document dates) and thus allow allowing the assessment of the relevance of a concept. |

## 4.4   ISA 500 – Scan other information available

*Background: Objective of ISA 500*

Most of the auditor's work in forming the auditor's opinion consists of obtaining and evaluating audit evidence from audit procedures performed during the course of the audit. However, audit evidence may also include information obtained from other sources inside and outside the entity. Audit evidence comprises both information that supports and corroborates management's assertions, and any information that contradicts such assertions. In addition, in some cases, the absence of information is used by the auditor, which constitutes audit evidence (ISA 500 para. 4, ISA 500 para. A1).

*Application: Practical issues*

While most of today's company data is qualitative in nature, e.g., websites, emails, professional publications and social networks, the audit currently focuses on analysing financial data. In fulfilling the requirements of ISA 500, the auditor primarily relies on inquiries with the entity's employees and investigating information available online, e.g., the entity's website.

An approach to increase efficiency in gathering, analysing and evaluating that information would not only increase the overall efficiency of the audit but also can lead to competitive advantages. Knowing things earlier is key in an information business.

*Application of the MOVING platform*

A tool to enable the efficient audit of information obtained from other sources inside and outside the entity, as required by ISA 500, should cover various sources of qualitative information available online, including the entity's website, newspaper websites and archives, and literature databases (i.e., recent publications). Thus, the auditor can quickly scan those sources of information for the entities name, related persons and user defined keywords to meet the requirements of ISA 500 efficiently. In addition to the requirements described in the previous sections, the following features are relevant to the ISA 500 scenario:

**Table 11:** ISA 500 requirements.

| ID | Topic | Keyword | Requirement |
|---|---|---|---|
| **#EY043** | Visualisation - top concepts | General | Two bar charts (one for entities, one for other concepts) that show the concepts that best describe the data. The length of each concept's bar depends on the relevance of the concept. |
| **#EY044** | Visualisation - top concepts | Sensitivity | There should be an option to change the amount of bars (i.e., concepts) displayed in each bar chart. |
| **#EY045** | Visualisation - top sources | General | A bar chart that shows the most relevant data sources for the given search query. The length of each concept's bar depends on the relevance of the source. |
| **#EY046** | Visualisation - top sources | Sensitivity | There should be an option to change the amount of bars (i.e., sources) displayed in the bar chart. |
| **#EY047** | Adaptive training support | User behaviour | The Adaptive Training Support (ATS) should provide the user with notifications about the navigation of other users through the system, e.g., which visualisation has been used, the most, with similar search settings or where other users went into the details (e.g., the expanding the network graph or going into the detailed data). |
| **#EY048** | Adaptive training support | Usage of search profiles | The ATS should include information about the number of search requires for an entity or identified key topics have been performed in the past. This can be combined with a recommendation to the user, for example, that other users who searched for entity A also searched for person B. |

## 4.5   ISA 520 – Inform on laws and regulations and changes therein

*Background: Objective of ISA 520 and effect of laws and regulations on the audit*

The effect on financial statements of laws and regulations varies considerably. Those laws and regulations to which an entity is subject constitute the legal and regulatory framework. The provisions of some laws or regulations have a direct effect on the financial statements in that they determine the reported amounts and disclosures in an entity's financial statements. Other laws or

regulations are to be complied with by management or set the provisions under which the entity is allowed to conduct its business but do not have a direct effect on an entity's financial statements. Some entities operate in heavily regulated industries (such as banks and chemical companies). Others are subject only to the many laws and regulations that relate generally to the operating aspects of the business (such as those related to occupational safety and health, and equal employment opportunity). Non-compliance with laws and regulations may result in fines, litigation or other consequences for the entity that may have a material effect on the financial statements (ISA 520 para. 2).

It is the responsibility of management, with the oversight of those charged with governance, to ensure that the entity's operations are conducted in accordance with the provisions of laws and regulations. This includes compliance with the provisions of laws and regulations that determine the reported amounts and disclosures in an entity's financial statements (ISA 520 para. 3).

The objectives of the auditor are:

- To obtain sufficient appropriate audit evidence regarding compliance with the provisions of laws and regulations generally recognised to have a direct effect on the determination of material amounts and disclosures in the financial statements.
- To perform specified audit procedures to help identify instances of non-compliance with other laws and regulations that may have a material effect on the financial statements.
- To respond appropriately to non-compliance or suspected non-compliance with laws and regulations identified during the audit (ISA 520 para. 10).

*Application: Practical issues*

Because of the high complexity of the numerous existing laws and regulations as well as its relationships, in the current audit practice there is no general solution for considering laws and regulations that may be relevant for the entity under audit. As every auditor has to evaluate which laws and regulations apply for every engagement he is working on, this is highly inefficient and bears the risk of omitting applicable laws and regulations. Furthermore, the auditor may not be up to date about recent changes in laws and regulations that may have a direct or indirect effect on the financial statements as required by ISA 520.

*Application of the MOVING platform*

The current practical issues in fulfilling the objectives of ISA 520 lead to the following requirements in addition to those described in the previous sections:

**Table 12:** ISA 520 requirements.

| ID | Topic | Keyword | Requirement |
|---|---|---|---|
| **#EY049** | Faceted search | Industries | Limit the search to certain industries, when applicable (e.g., retail, automotive, airlines). This can support the auditor in identifying the relevant laws |

| | | | and regulations. |
|---|---|---|---|
| **#EY050** | Entity identification | Abbreviations | The concept identification algorithm should recognise abbreviations and treat the abbreviated and the written-out word as one concept. This especially applies to laws and regulations (e.g., "HGB" and "Handelsgesetzbuch"). |
| **#EY051** | Search list visualisation | General | A conventional search result list. By default, the relevance is the sorting criteria for results. The user can choose to sort by other criteria such as name, author or publication date. |
| **#EY052** | Adaptive training support | Search list | When searching for specific paragraphs or laws and regulations, the ATS should remind the user that it might be useful to sort the search list by publication date in order to address recent changes and comments that might be relevant. |
| **#EY053** | Adaptive training support | Search term | When searching for a specific law or a regulation, the ATS should notify the user about the laws and regulations other users (who also search for that law) searched for. |
| **#EY054** | Visualisations - general | Filter by geographical area | Based on the search results, a world map displays any geographic data contained by markers on the map. The user can limit the search results to specific locations by selecting and unselecting them. |

## 4.6   ISA 230 – Quality of audit documentation

*Background: Objective of ISA 230*

The objective of the auditor is to implement quality control procedures at the engagement level that provide the auditor reasonable assurance that: (a) the audit complies with professional standards and applicable legal and regulatory requirements; and (b) the auditor's report issued is appropriate in the circumstances (ISA 230 para. 6). Key element of any internal or external audit quality control procedures is the audit documentation prepared by the engagement team. According to ISA 230 para. 5, the auditor is required to prepare documentation that provides: (a) a sufficient and appropriate record of the basis for the auditor's report and (b) evidence that the audit was planned and performed in accordance with ISAs and applicable legal and regulatory requirements.

*Application: Practical issues*

To ensure that the audit documentation is in accordance to the requirements of ISA 230, engagement executives need to review the work performed, the issues identified and the conclusions reached by the engagement team. In an average statutory audit of the financial statements, there are between 300 and 500 documents to review. These documents are of different file formats and do, in most cases, not have a standardised document name. In addition to review these large amounts of unstructured data, it is currently highly inefficient for the engagement executives to search for a specific information within the data, so the audit must ensure addressing the topic and covering it in documentation.

*Application of the MOVING platform*

The current practical issues in fulfilling the objectives of ISA 230 lead to the following requirements in addition to those described in the previous sections.

**Table 13:** ISA 230 requirements.

| ID | Topic | Keyword | Requirement |
|---|---|---|---|
| **#EY055** | Faceted search | Data sources | We should consider enabling the connection of the MOVING platform to files stored in the intranet. However, the tool may not upload those files to the internet at any stage of the analysis. |
| **#EY056** | Visualisations - general | Search feature | There should be a search function to locate certain keywords within the visualisations. |
| **#EY057** | Visualisations - general | Drill down | At any stage of the analysis, the user needs to drill down to the detailed data. When selecting a concept (e.g., a tag in the tag cloud or a node in the network graph), the user should be able to receive a list of the corresponding documents, including the links to open these documents. |
| **#EY058** | Visualisations - general | Export | To facilitate the review from engagement executives, the user should be able to export the visualisations. In the export, the work done within the program should be documented, i.e., the search settings and the steps performed to adjust the visualisation. |

## 4.7   ISA 240 – Search for unusual journal entry descriptions

*Background: Objective of ISA 240*

ISA 240 requires the auditor to: (a) identify and assess the risks of material misstatement of the financial statements due to fraud, (b) obtain sufficient appropriate audit evidence regarding the assessed risks of material misstatement due to fraud, through designing and implementing appropriate responses; and (c) respond appropriately to fraud or suspected fraud identified during the audit (ISA 240.10). Material misstatement of financial statements due to fraud often involves the manipulation of the financial reporting process by recording inappropriate or unauthorised journal entries. The auditor shall design and perform audit procedures to test the appropriateness of journal entries recorded in the general ledger (ISA 240.32(a)).

*Application: Practical issues*

The audit procedures to test the appropriateness of journal entries include, for example, selecting journal entries posted by unauthorised persons, journal entries made at the end of the reporting period or on weekends or journal entries that contain unusual number patterns. While existing audit programs can automise those tasks, the analysis of journal entries containing unusual booking descriptions remains a challenge in the auditor's daily business.

The current approach for analysing journal entry descriptions in accordance with ISA 240 is to filter the data using defined criteria, e.g., names of persons that are not authorised to post journal entries or key words that indicate the journal entry was posted on somebody else's behalf. However, this manual search is highly inefficient and limited by the creativity of the auditor and his experience with the entity. Moreover, it does not take into account potential orthographic mistakes or the fact that a potential fraudster could choose a usual journal entry description for an unusual transaction.

*Application of the MOVING platform*

Inverting the current approach to audit qualitative journal entry data can contribute to meet the requirements of ISA 240 in an efficient manner. The auditor should apply text-mining techniques to investigate all journal entry descriptions in a structured manner instead of filtering for previously defined criteria.

The current practical issues in fulfilling the objectives of ISA 240 lead to the following requirements in addition to those described in the previous sections:

**Table 14:** ISA 240 requirements.

| ID | Topic | Keyword | Requirement |
|---|---|---|---|
| **#EY060** | Entity identification | Uniqueness | As the entity-identification is key to the EY use cases, we should consider including "dictionaries" from DUNS (Data Universal Numbering System for the unique identification of entities) to take |

| | | | advantage of the uniqueness of those systems. |
|---|---|---|---|
| **#EY061** | Entity identification | Excel files | As journal entry descriptions usually contain multiple words, the program should be able to separate multiple words stored in MS Excel cells. |
| **#EY062** | Entity identification | Abbreviations | As there can be several abbreviations contained in the data, the tool should contain a dictionary to recognise common abbreviations. |
| **#EY063** | Date mentions | General | This visualisation does not show the data but a calendar. The font size of the days in the calendar depends of the frequency those dates are mentioned within the data. When a day is not mentioned within the data, it is not visible in the visualisation. |
| **#EY064** | Date mentions | Sensitivity | The user can choose to display only dates mentioned more than X times within the data. |
| **#EY065** | Visualisations - general | Tagging | The user should be able to tag/mark concepts for further analysis and thus be able to export a summary containing the tagged concepts and (a list of) the underlying data sources. |

# 5 Use case requirements use case 2: managing and mining research information

The following subsections describe the requirements of use case 2 that are aligned to the framework of reference for information literacy (Klingenberg, 2016a, Klingenberg, 2016b). Therefore, super epics were formulated describing comprehensive working processes the user conducts on the platform. Super epics consist of several epics that define sub-steps of every working process. For example, it is assumed, that the super epic "searching" as comprehensive working process consists of the epics (sub-steps) "express the need for further knowledge", "conduct query", "refine search results", "verify completeness of the query" and "select and look at several search results". Then the user requirements for every sub-step, derived from the interviews and formulated as user stories (see section 3.2 for the respective methodology), are presented in tabular form. The several steps for information and knowledge investigation are related to the scenarios 1 to 3 of the second use case. The requirements for scenario 4 will be presented in D2.1 (Section 3).

## 5.1 Super Epic: search on the MOVING platform

Searching in this context contains the whole searching-process, from expressing a query, refining the list of search results by verifying what is relevant and what is not, to segregating information and documenting the searching process. "Searching" is therefore a Super Epic, because it can be dissected into different epics.

**Express the need for further knowledge:** The searching process starts with the expression of a query, i.e. the expression of a need for further knowledge to answer the question "What do I want to know?" (Klingenberg, 2016a). The user thereby first decides whether he/she wants to conduct a state of the art (Use Case Scenario 1) or search for project partners (Use Case Scenario 2) or search for funding opportunities (Use Case Scenario 3).

**Conduct query:** After the expression of a need for knowledge, the user starts the query to find resources. Conducting the query is driven by the question: Where could I find it?

**Refine search results:** After having a long list with search results, the user refines his/her search results by verifying the relevance of the results to his research topic, i.e. chooses resources that are relevant to his research topic (i.e.: Where is it written?).

**Verify completeness of the query:** After refining the search results, the user verifies the completeness of the query and decides, whether the remaining search results completely answer the (re)search question (i.e.: Is that all?), or further queries are necessary.

**Select and look at several search results:** The user selects several search results to take a closer look at them, e.g. to read them and then decide what to do with them.

**Verify factual accuracy of several search results:** The user verifies the factual accuracy of several search results by comparing different content sources. Then he/she decides, if one specific search result can be further used or not.

**Document the search process:** The user documents the search process by capturing used search terms and search results, writing down which search terms he/she included or excluded.

### 5.1.1 Epic: express the need for further knowledge

**Table 15**: Express the need for further knowledge.

| ID | Role | User Story |
|---|---|---|
| **#TUD001** | USER | I want to search for combination of words. |
| **#TUD002** | USER | I want to include or exclude search terms. |
| **#TUD003** | USER | I want to see which words/search frequently occur with each other (thesauri, synonyms). |
| **#TUD004** | USER | I want to set up my own search profile for searching of documents on the platform. |
| **#TUD005** | USER | I want to see whether one search term or keyword has different meanings. |
| **#TUD006** | USER | I want to search for essential terms or concepts of a research topic to find relevant literature. |
| **#TUD007** | USER | I want to search for possible partners for projects in other countries. |
| **#TUD008** | USER | I want to look for possible business partners. |
| **#TUD009** | USER | I want to inform myself about other researchers. |
| **#TUD010** | USER | I want to see, if authors occur frequently during my research. |
| **#TUD011** | USER | I want to find a specific historical source. |
| **#TUD012** | USER | I want to find digitised journals in the humanities. |

### 5.1.2 Epic: conduct query

**Table 16:** Conduct query.

| ID | Role | User Story |
|---|---|---|
|  |  |  |

| ID | Role | User Story |
|---|---|---|
| **#TUD014** | USER | I want to do a full-text search on the platform. |
| **#TUD015** | USER | I want to search for the same search term within different sources e.g. different databases, social media, Websites to get information about a new research topic. |
| **#TUD016** | USER | I search within databases to find historical information in my discipline (history), e.g. of the Bundesarchiv. |
| **#TUD017** | USER | I want to search in social media to verify the social relevance of a specific topic. |
| **#TUD018** | USER | I want to search on social media (e.g. Social Networks, Twitter) to become informed about new papers and publications. |
| **#TUD019** | USER | I want to search in social media to get information about past scientific events. |
| **#TUD020** | USER | I want to search in social media and on Websites to find up-to-date information on the topic I want to examine, which is available in the literature only months later. |
| **#TUD021** | USER | I want to search within social media to get current scientific information and to look over the edge of my discipline. |
| **#TUD022** | USER | I want to search in existing scientific social networks, e.g. research gate, academia.eu for other researchers. |
| **#TUD023** | USER | I want to search within existing social networks like XING, LinkedIn. |
| **#TUD024** | USER | I want to search within blogs and on specific websites, because I want to get an overview over a broader topic (e.g. agile project management). |
| **#TUD025** | USER | I want to search in Wikipedia and I watch YouTube-Videos to get first information on a specific topic (a first insight) and to researching initial information on a specific question from a different topic. |

| ID | Role | User Story |
|---|---|---|
| **#TUD026** | USER | I want to get information from Wikipedia, Google, Yahoo, Bing etc. |
| **#TUD027** | USER | I want to search within existing library catalogues. |
| **#TUD028** | USER | I search in the catalogue of my library in order to obtain publications, specialised articles, books and meta-information on publications (in particular page numbers, annual figures, keywords, classification data). |
| **#TUD029** | USER | I want to search within plenary protocols of the German Bundestag[81]. |
| **#TUD030** | USER | I use the Website of the Perseus-Project to find antique texts, which are original or translated into German or English. |
| **#TUD031** | USER | I use archive.org to look at older websites that are already offline. |
| **#TUD032** | USER | I need to get an overview over databases that are relevant to my research field. |
| **#TUD033** | USER | I want to ask experienced researchers a question. |

### 5.1.3 Epic: refine search results

**Table 17:** Refine search results.

| ID | Role | User Story |
|---|---|---|
| **#TUD034** | USER | I want to research on legal texts. |
| **#TUD035** | USER | I want to search for scientific literature, i.e. books or journal articles. |
| **#TUD036** | USER | I want to search for videos. |
| **#TUD037** | USER | I want to search for survey data. |
| **#TUD038** | USER | I want to filter articles by date of publication to get the most current publications. |

---

[81] http://www.bundestag.de/en/documents/minutes_neu (2017-03-27)

| #TUD039 | USER | I want to filter my search results by title, abstract or full-text. |
|---|---|---|
| #TUD040 | USER | I want to filter articles by author to find out, if the publication is relevant to me. |
| #TUD041 | USER | I want to filter my search results by discipline. |
| #TUD042 | USER | I want to filter the search results due to the amount of the citations. |
| #TUD044 | USER | I want to filter search results by topic. |
| #TUD045 | USER | From my list with search results, I want to pick those search results out, relevant for my research topic. |
| #TUD046 | USER | I want to filter search results by the number of clicks of other MOVING users. |
| #TUD047 | USER | I want to see in my list of search results which publications are frequently cited in other publications (snowball method) in order to get involved in a new topic. |
| #TUD048 | USER | I want to see the deadline and the contact of a funding opportunity, to decide whether it fits to my research project or not. |
| #TUD049 | USER | I want to be able to verify through which resources I have already searched. |
| #TUD050 | USER | I want to detect relevant disciplines and search term occurrence, so that I can get search results that go beyond the search term, I am looking for. |
| #TUD051 | USER | I want to know how often an author is cited in other publications, relevant to my research topic, because for me, this is an indicator for their importance within the research field. |
| #TUD053 | USER | I do not want to have duplicates in my search list when cross searching in different databases. |
| #TUD054 | USER | I want to see the frequency of usage tendencies of keywords (frequency of keyword usage) used for tagging of several search results, to decide whether those search results fit to my research topic or not. |
| #TUD055 | USER | I want to get an overview over patterns, concepts, people and phrases recurring in my search results to get a deeper insight into a topic. |
| #TUD056 | USER | I want to choose options for possible visualisation of search results. |

| #TUD057 | USER | I want to visualise my search results as topic network, to see which results are relevant to my research topic. |
| #TUD058 | USER | I want to visualise my search results as tag-cloud of topics, to see which results are relevant to my research topic. |

### 5.1.4 Epic: verify completeness of the query

**Table 18:** Verify completeness of the query.

| ID | Role | User Story |
|---|---|---|
| **#TUD059** | USER | I want to have a checkbox next to a resource so I can indicate, whether I already had a look at this reference. |
| **#TUD060** | USER | I want get a list with references (search results) displayed to get an overview. |
| **#TUD061** | USER | I want to visualise my search results as network, which I can move by mouse-click and where I can zoom in and out, to be able to set an individual focus and get a broader overview over my search results. |
| **#TUD062** | USER | I want to have visualisation of search results, that shows me for example, if I search for the term X and "Mueller et al." has been contained in all sources, that I then can visualise which sources are included in all found papers, to get an comprehensive overview over relevant literature. |
| **#TUD063** | USER | I want to visualise my search results as tag-cloud to keep track of used keywords. |
| **#TUD064** | USER | I want to visualise my search results as authors' network. |

### 5.1.5 Epic: select and look at several search results

**Table 19:** Select and look at several search results.

| ID | Role | User Story |
|---|---|---|
| **#TUD065** | USER | I want to select several search results from my list of search results. |

| #TUD066 | USER | I want to open several search results in different tabs, to have a look at them. |
|---------|------|----------------------------------------------------------------------------------|
| #TUD067 | USER | I want to save several search results as bookmarks, to return to a specific source easily. |
| #TUD068 | USER | I want to search for relevant keywords within a text for further searching on my research topic. |
| #TUD069 | USER | I want to be able to "jump" from one word or section in a text to another section or word in that text when I click on a specific word within the text. |
| #TUD070 | USER | I want to "jump" from one text to another when I click on a specific Keyword within the text to find further important resources for my research. |
| #TUD071 | USER | I want to know where (e.g. in which library) I can find book from my search results, when I click on it. |
| #TUD072 | USER | I want to mark high and low relevance of search results (e.g. literature, journal paper) with a colour regarding to my research topic. |
| #TUD073 | USER | I want to read specific texts of my search results on the platform to find topic-related keywords within that text. |
| #TUD074 | USER | I want to have an OCR (e.g. for scanned texts), so I can mark text sequences and search texts. |

### 5.1.6 Epic: verify factual accuracy of several search results

**Table 20:** Verify factual accuracy of several search results.

| ID | Role | User Story |
|----|------|------------|
| #TUD075 | USER | I compare Wikipedia entries in different languages, to verify their factual accuracy. |
| #TUD076 | USER | I look up the line of argumentation in discussion forums on a specific topic (e.g. discussion page of a specific Wikipedia article) and the sources in footnotes to examine the factual accuracy of an article. |
| #TUD077 | USER | I compare information from different resources (e.g. Wikipedia articles with literature) on my research topic to verify the seriousness of the information. |

### 5.1.7 Epic: document the search process

**Table 21:** Document the search process.

| ID | Role | User Story |
|---|---|---|
| **#TUD078** | USER | I need a functionality that allows me to keep track to my search terms, search results and databases I searched within to be able to exclude search terms for further searching, while I am doing a structured literature review. |
| **#TUD079** | USER | I need a meta data library/ meta data management to tag my search results with keywords. |
| **#TUD080** | USER | I want to write down which search terms I used to search within the data bases and how I used the Boolean operation, i.e. what search terms I excluded and what search terms I could possibly additionally use with 'AND'. |
| **#TUD081** | USER | I want to save my faceted search configurations. |
| **#TUD082** | USER | I want to tag my queries so I can find them easily later on. |
| **#TUD083** | USER | I want to have a systematic reference management to keep track of my search results (e.g. literature, journal articles) and to not miss resources or have them twice included. |

## 5.2 Super Epic: generate knowledge out of information

After researchers searched for information in the MOVING platform and selected several search results, they generate knowledge by dealing with the selected search results. They formulate own thoughts based on several search results; they compare different information and formulate answers on research questions. Additionally, several search results are considered in a broader context and their importance for the particular research topic is evaluated. Part of the generation of knowledge is also the organisation of information with self-developed criteria (Klingenberg, 2016a).

**Formulating knowledge**: researchers express selected information with own words by writing notes in the margin of texts or writing excerpts of full texts.

**Compare information with other information**: researchers compare information from several search results; they adopt already known information and new information onto their research topic.

**Integrate information into a given context:** researchers appraise the meaning of an information for the research topic and develop or rethink answers on their research questions with the help of this information.

**Organise - connect several information in a meaningful way**: researchers approach several search results with the help of a given structure, dissemble texts or other content into meaningful entities and approach it with the help of keywords, which they mark by themselves.

### 5.2.1 Epic: formulating knowledge

**Table 22:** Formulating knowledge.

| ID | Role | User Story |
|---|---|---|
| **#TUD084** | USER | I want to be able to make notes of or excerpt the texts I read, to summarise the content, get a first overview and internalise what I have read. |
| **#TUD085** | USER | I want to organise and administrate my notes or the excerpts I wrote, to be able to evaluate my literature research. |
| **#TUD086** | USER | I want to be able to copy and paste the notes I wrote on MOVING directly into my own publications and to download my notes on my PC. |

### 5.2.2 Epic: compare information with other information

**Table 23:** Compare information with other information.

| ID | Role | User Story |
|---|---|---|
| **#TUD087** | USER | I want to compare the content of different search results e.g., I want to compare the content of books with the content of journal articles. |
| **#TUD088** | USER | I want to compare funding options. |
| **#TUD089** | USER | I want to search within different databases and compare the search results from the different databases. |
| **#TUD090** | USER | I want to see connections between search results from journals, archives, conferences, books when conducting a search on one specific topic. |

### 5.2.3    Epic: integrate information into a given context

**Table 24:** Integrate information into a given context.

| ID | Role | User Story |
|---|---|---|
| **#TUD091** | USER | I want to visualise different terms which have a contextual connection to each other as word-web, to identify connections between terms and to see which terms are used within the same context ("with ties that show what belongs together i.e. if it is talked about one term often it is talked about another term, too."). |
| **#TUD092** | USER | I want to have an automatised tree structure-visualisation of search results (e.g. journal paper) that build upon each other to identify relations between journal papers and then be able to deviate how to go on working through/ reading them. |
| **#TUD093** | USER | I search for keywords and their occurrence in specific periods in different successive documents and I can imagine a visualisation of search results as timeline to show their chronological appearance. |
| **#TUD094** | USER | I want to look at the temporal course of the public discourse on my research topic, to investigate how often and when a certain term, relating to that topic, was used. |

### 5.2.4    Epic: organise - connect several information in a meaningful way

**Table 25:** Organise - connect several information in a meaningful way.

| ID | Role | User Story |
|---|---|---|
| **#TUD095** | USER | I want to be able to save my search results on the MOVING platform, so I can find them later. |
| **#TUD096** | USER | I want to organise the found resources directly on the platform, e.g. through a reference system. |
| **#TUD097** | USER | I want to be able to tag search results by myself with keywords I defined, to be able to categorise and handle them. |
| **#TUD098** | USER | I want to be able to export my search results into my reference manager (e.g. Zotero, Mendeley). |

| #TUD099 | USER | I want to have a reliable meta-data library, so I do not have to check all meta data of my literature saved. |
|---|---|---|
| #TUD100 | USER | I want to have the search terms I used marked with a colour within the text of the search results. |
| #TUD102 | USER | I want to have access to full-texts so I can read and annotate them and mark keywords within them. |
| #TUD103 | USER | I want to make annotations and mark words or sections in texts, relevant for my own research publications. |
| #TUD104 | USER | I want to have a reference management-functionality to systematically organise and manage my search results. |
| #TUD105 | USER | I want to tag words, phrases or sections within texts with keywords by myself on the MOVING platform to find specific terms again. |
| #TUD106 | USER | I want to mark the found literature with colours to organise them. |
| #TUD107 | USER | I want to have access to the MOVING platform from different devices. |
| #TUD108 | USER | I want to make annotations on specific sequences in videos to be able to analyse them. |
| #TUD109 | USER | I want to download the journal articles or other search results, which are relevant for my research, on my own PC to be able to organise the literature I have found. |

## 5.3   Super Epic: community build-up

**Sharing**: researchers share resources and information with other researchers. To make citations of resources in publications for example is one possibility to share information within the scientific community. Other possibilities are the usage of social networks or other social media and community tools to share resources like texts, links, videos or just thoughts. For both kinds of sharing, it is necessary for researchers to clarify the terms of use of resources, before using it for citations or other kinds of sharing.

**Communication and collaborative work**: besides the sharing of resources, researchers also connect and communicate with each other e.g. they write emails, build-up their own personal network or work collaboratively.

### 5.3.1 Epic: sharing

**Table 26:** Sharing.

| ID | Role | User Story |
|---|---|---|
| **#TUD110** | USER | If I upload a paper on the MOVING platform by myself, I want to have a clear overview, whom of the platform users is allowed to see my uploaded articles and I want to be able to mark whether it is OA or not. |
| **#TUD111** | USER | I want to double check information from Wikipedia articles with scientific literature on the same topic to make the content found in Wikipedia scientifically quotable. |
| **#TUD112** | USER | I want to see whether the resource is open access or not. |
| **#TUD113** | USER | I want to use creative commons resources for my research (e.g. from Flickr). |
| **#TUD114** | USER | I need plugins (BibTex) to create literature references to integrate them, automatically, into my own texts. |
| **#TUD115** | USER | I want to share pictures I have found on the MOVING platform with other users. |
| **#TUD116** | USER | I want to share literature I have found on the MOVING platform with my colleagues and fellow students |
| **#TUD117** | USER | I want to rank the found resources so for other users it is possible to see, whether the search results fitted to my query well or not so well |
| **#TUD118** | USER | I want to be able to see which content is Open Access or Creative Commons, because otherwise I cannot find Open-Access-Publications on the platform. |

### 5.3.2 Epic: communication and collaborative work

**Table 27:** Communication and collaborative work.

| ID | Role | User Story |
|---|---|---|
| **#TUD119** | USER | I want to create collaborative texts on the platform/to work collaboratively on the platform. |

| #TUD120 | USER | I want to work collaboratively on the platform. |
|---|---|---|
| #TUD121 | USER | I want to ask experienced researchers a question. |
| #TUD122 | USER | I want to get proposals for search queries from other users (What did they do at that specific step of search?). |
| #TUD123 | USER | I want to contact myself about researchers on existing scientific social networks, e.g. research gate, academia.eu. |
| #TUD124 | USER | I want to see the profile of the author, which I am looking for, in order to be informed about the publication/quotations/cross-references. |
| #TUD125 | USER | I want to know the contact details of the authors from journal articles. |
| #TUD126 | USER | I want the contact details from frequently occurring authors to be displayed. |
| #TUD127 | USER | I use the Wikipedia resource exchange to find journal papers I cannot access through the library university. |
| #TUD128 | USER | I contact libraries to get access to books that I cannot access via my university. |

# 6    Functional requirements

This section describes which functions the MOVING platform should have due to above the described requirements for both use cases (Section 3). The table shows first the ID of the requirement, which is named in the next column, and describes the specific topic of the requirement. The keyword column describes which detail of the topic is addressed. The next column is for the description of the topic. The reference to the use case IDs shows, which requirements of the use cases are addressed with the described requirement to the platform.

## 6.1    Requirement: search field

**Table 28:** Search field.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#001** | Author | | There should be search field where to search for an author, a title or a key word when searching for resources. | #TUD040 |
| **#002** | Title | | | |
| **#003** | Key word | | | #TUD006, #TUD015 |
| **#004** | Combination of words | | It should be possible not only to search for one word but for combination of words. | #TUD001 |
| **#005** | Including and excluding words | | It should be possible to exclude and include words or terms, e.g. when the user is unsure about a word or a search term, he/she can enter the part of the search term he is sure about and add ** behind it to indicate, whether there is a part of the search term missing he/she does not know. | #TUD002 |

## 6.2 Requirement: faceted search

**Table 29:** Faceted search.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| #006 | Date | | Refining the search after the publishing date of the resource. Specify the relevant date range (e.g., the period after the entity's formation). | #TUD038, #EY036 |
| #007 | Search term | | Exclude one or multiple search terms. | #EY001 |
| #008 | | | Define one or multiple search terms. | #EY002 |
| #009 | Type of content | | Searching in and possibility of excluding monographies, journal articles, open access journal articles, conference articles, posters. | #TUD035 |
| #010 | Search extent | Search depth | Refine the search by title, abstract, full-text. | #TUD039, #EY008 |
| #011 | Databases | | Searching and possibility of including or excluding databases (see annex no. 9.4). Connection to various literature databases available to include publications of several disciplines. | #TUD011, #TUD012, #TUD016, #TUD089, #EY007 |
| #012 | | | Listing of all databases (sorted by discipline). | #TUD032 |
| #013 | Data sources | | Include or exclude certain data sources. | #EY003 |
| #014 | | World Wide Web | Data sources should include both the world wide web and the possibility to define one or multiple specific websites to include. | #EY004 |
| #015 | | Various extension types | Browse for files stored on the local device to include in the analysis. Those files can be of various file formats, e.g., .pdf, .doc, | #EY005 |

| | | | .doc, .rtf, .txt, .xls, .xlsx, .csv, .htm, .html. | |
|---|---|---|---|---|
| #016 | | Data source upload | The upload of data sources stored on the local device is restricted due to data privacy issues. | #EY006 |
| #017 | | Intranet | We should consider enabling the connection of the MOVING platform to files stored in the intranet. However, the tool may not upload those files to the internet at any stage of the analysis. | #EY055 |
| #018 | Auto-complete | | A recommendation/auto-completion feature supports the specification of the search term(s). This will support the auditor in: (a) searching for the correct entity and (b) identifying similar (and therefore possibly related) entities right from the beginning of the analysis. | #EY034 |
| #019 | Dynamic Interface | | Based on the search settings, the remaining search criteria become enables or disabled (e.g., when a file on the local device is included that has no metadata, the depth of the search cannot be set to metadata). | #EY011 |
| #020 | Geographic Region | Location search | Limit the search to certain geographical areas (e.g., the entity's headquarters country). | #EY009 |
| #021 | Language | | Limit the search to specific languages. By default, the language is in correspondence to the search term defined. | #EY012 |
| #022 | Concepts | Concept Type | Ability to limit the search to certain types of concepts, for example, persons and organisations only. | #EY037 |
| #023 | Industry | Industry Type | Limit the search to certain industries, when applicable (e.g., retail, automotive, airlines). This can support the auditor in | #EY049 |

| | | | | |
|---|---|---|---|---|
| | | | identifying the relevant laws and regulations. | |
| #023 | Author | | Refine the search by author by excluding and including authors. | #TUD040 |
| #024 | Survey data | | Searching and possibility of including or excluding survey data (GESIS, Statista[82]). | #TUD037 |
| #025 | Discipline | | Refining the search due to including or excluding scientific disciplines. Limit the search to certain disciplines, when applicable (e.g., when selecting a database that covers multiple disciplines). | #TUD041, #TUD050, #EY010 |
| #026 | Citations | | Refine the search after the amount of citations of the source. | #TUD042, #TUD051 |
| #027 | Clicks | | Refine the search after the amount of clicks on the source through the users. | #TUD046 |
| #028 | User profile | | It should be possible, that the user can adjust the faceted search due to his/her needs. E.g., the specific user searches mostly within the social sciences, so he/she saves this for the faceted search. | #TUD004, #TUD081 |
| #029 | Social media | | Searching in and possibility of including or excluding social media (Twitter, Facebook, ResearchGate, academia.eu, XING, LinkedIn, Blogs). | #TUD017, #TUD018, #TUD019, #TUD020, #TUD021, #TUD022, #TUD024, #TUD123, #EY035 |
| #030 | Websites | | Searching in and possibility of including and excluding websites (see annex no. | #TUD020, #TUD024, |

---

[82] https://de.statista.com (2017-03-27)

| | | | 9.6). | #TUD030, #TUD034 |
|---|---|---|---|---|
| **#031** | Online Encyclopaedia | Wikipedia | Searching in a possibility of including and excluding Wikipedia. | #TUD025, #TUD026 |
| **#032** | | | Possibility of comparing Wikipedia articles from different languages. | #TUD075 |
| **#033** | Website | archive.org | Searching in a possibility of including and excluding archive.org. | #TUD031 |
| **#034** | Search engines | Google, Bing, Yahoo | Searching and possibility of including or excluding existing search engines. | #TUD026 |
| **#035** | Library catalogues | | Searching and possibility of including or excluding library catalogues (see annex no. 9.5). | #TUD027, #TUD028 |
| **#036** | PDF files | e.g. Plenary protocols of the Bundestag | Searching and possibility of including or excluding PDF files. | #TUD029 |
| **#037** | Project partners | international | It should be possible to look for possible project partners from different countries. | #TUD007 |
| **#038** | Business partners | international | It should be possible to look for business partners via the platform. | #TUD008, #TUD023 |
| **#039** | Videos | | Searching and possibility of including or excluding videos. | #TUD036 |
| **#040** | Data source for funding | | Possibility of including or excluding funding databases. | #TUD088 |
| **#041** | Creative commons | Flickr | Searching in and possibility of excluding or including sources under the licence of creative commons (e.g. Flickr[83]). | #TUD113 |

---

[83] https://www.flickr.com (2017-03-27)

## 6.3   Requirement: search list

**Table 30:** Search list.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#042** | List | | Displaying the search results through a list. | #TUD060, #EY051 |
| **#043** | Selecting | | Picking up the relevant search results through clicking on it. | #TUD045, #TUD065 |
| **#044** | Display frequently cited reference per source | | Displaying frequently cited references per source through mouse over. | #TUD047 |
| **#045** | Funding deadline | | Displaying the deadline for a funding opportunity and possibility of ordering the results due to the deadline. | #TUD048 |
| **#046** | Funding contact | | Displaying the contact for a funding opportunity by mouse over. | #TUD048 |
| **#047** | Duplicates | | Avoiding duplicates due to search in different databases. | #TUD053 |
| **#048** | Checkbox | | Displaying which resources, I already searched/looked at through a checkbox. | #TUD049, #TUD059 |
| **#049** | Different tabs | | Open search results in different tabs. | #TUD066 |
| **#050** | Bookmark | | Bookmark several search results. | #TUD067 |
| **#051** | Library connection | | Displaying where the source can be found (e.g. in a library) by clicking on it. | #TUD071, #TUD128 |
| **#052** | Mark results | | Marking the results due to high or low relevance on the search topic. | #TUD072 |
| **#053** | Open access | | Displaying whether the source is open access or not. | #TUD112, #TUD118 |

| #054 | Creative commons | | Displaying whether the source is under the licence of creative commons or not. | #TUD113, #TUD118 |
| #055 | Author Contact | | Displaying the contact details of the author of the resource (especially from frequently occurring authors). | #TUD125, #TUD126 |

## 6.4   Requirement: visualisation

**Table 31:** Visualisation.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| #056 | Different meanings of search term | | It should be possible to see via the visualisation whether the search term/word has different meanings due to different disciplines. | #TUD005 |
| #057 | connections between different types of sources | | It should be possible to see connections between different types of sources, e.g. journals, archives, conferences, books. | #TUD090 |
| #058 | Different types of visualisation | | Choosing different options of visualisation. | #TUD056 |
| #059 | Top Concepts | General | Two bar charts (one for entities, one for other concepts) that show the concepts that best describe the data. The length of each concept's bar depends on the relevance of the concept. | #EY043 |
| #060 | | Sensitivity | There should be an option to change the amount of bars (i.e., concepts) displayed in each bar chart. | #EY044 |
| #061 | Top sources | General | A bar chart that shows the most relevant data sources for the given search query. The length of each concept's bar depends | #EY045 |

| | | | on the relevance of the source. | |
|---|---|---|---|---|
| **#062** | | Sensitivity | There should be an option to change the amount of bars (i.e., sources) displayed in the bar chart. | #EY046 |
| **#063** | Topic network | Concept Graph | Visualisation as topic network. Concept graph to visualise topics around a specific keyword. The graph consists of notes and edges linking those nodes. The nodes are the most relevant concepts within the data and the edges link the concepts (nodes) that frequently co-occur within the data. | #TUD057, #TUD091, #EY017 |
| **#064** | | Filter | Ability to limit the view to certain types of concepts, for example, persons and organisations only. | #EY038 |
| **#065** | Network | Navigation | The user should be able to navigate through the network, diving deeper into areas of interest and expanding the network in relevant directions. To increase the efficiency of the analysis, there should be an indication for the user how many documents/entities/locations/concepts will become visible when expanding the network accordingly. | #EY018 |
| **#066** | | Completeness | There should be an indicator of the degree of completeness on the current view. This should be based on the additional information that can be obtained by extending the network. | #EY019 |
| **#067** | | Filter | There should be a general option to change the displayed amount of nodes and edges. With the minimum settings, the graph only shows the most relevant nodes/edges. | #EY020 |

| #068 | | Size of nodes, thickness of edges | By default, the size of the node depends on the concept's relevance (e.g., frequency of occurrence within the data) and the thickness of the edges depends on the number of co- occurrences of the connected nodes. We should consider adding alternatives, e.g., sizing the nodes by in-degree or out-degree. | #EY021 |
|---|---|---|---|---|
| #069 | | Co- occurrence of concepts | Ability to specify the co-occurrence of concepts that is displayed in the graph (e.g., concepts connected by edges are included in the same text, paragraph or sentence). | #EY022 |
| #070 | | Hovering | When hovering the mouse over nodes and edges, there should be additional information about the attributes of the node/edge (e.g., number of occurrences, data source with the most occurrences). | #EY023 |
| #071 | | Statistics and measures | The network development should go hand in hand with statistical measures, e.g. centrality, concentration, density, shortest path, community clustering. These measures will be determined in more detail in D1.3. | #EY024 |
| #072 | | Colour | The colour of the nodes corresponds to the "entity identification" algorithm, e.g., person = red, organisation = green, location = blue, other concept = black. | #EY039 |
| #073 | Topic network | | Moving the topic network by click and hold. | #TUD061, #EY031 |
| #074 | | | Zooming in and zooming out of the network. | #TUD061, #EY031 |
| #075 | Author network | | Visualisation of the authors of the sources. | #TUD064 |

| #076 | One network for all | | Visualisation of patterns, concepts, people and phrases in one network. | #TUD055 |
|---|---|---|---|---|
| #077 | Focus on references in the sources | | Visualisation of frequently occurring references in the search results. | #TUD062 |
| #078 | Tag cloud | General | Visualisation as topic as tag cloud of topics. The tag cloud displays the top keywords extracted from the data. | #TUD058, #TUD063, #EY025 |
| #079 | | Size | The initial size of the tag depends on the tag's frequency within the search results. | #EY026 |
| #080 | | Position | The position of the tag within the tag cloud depends on the relevance. The initial position of the tag within the tag cloud depends on the keyword's relevance. However, there should also be an option to order the tags alphabetically. | #EY027 |
| #081 | | Rating of tags | When reviewing the initial tag cloud, the user can "rate" the different tags according to his research question/area of interest. Rating tags leads to a refresh of the tag cloud taking into account these ratings. | #EY028 |
| #082 | | Sensitivity | There should be an option (e.g., a scroll bar) to adjust the sensitivity of the analysis. When changing the sensitivity, the size of the tag cloud changes as the required frequency/relevance of a concept to be included in the tag cloud changes. | #EY029 |
| #083 | | Colour | The colour of the tag corresponds to the "entity identification" algorithm, e.g., person = red, organisation = green, location = blue, other concept = black. | #EY040 |

| #084 | | Learning | The platform should learn from the user behaviour. For example, when a number of users rated a tag (e.g., a related organisation) as "high", the standard rating/weight for this tag in future search queries should adjust automatically. | #EY041 |
|---|---|---|---|---|
| #085 | Tag cloud for key words | | Visualisation of key words used in the databases and on the platform by users. | #TUD054 |
| #086 | Tagging | | The user should be able to tag/mark concepts for further analysis and thus be able to export a summary containing the tagged concepts and (a list of) the underlying data sources. | #EY065 |
| #087 | Tree visualisation | | Tree visualisation of search results to see which results are based on another. | #TUD092 |
| #088 | Timeline visualisation | | A timeline visualisation for seeing how the search results appear chronological. | #TUD093, #TUD094 |
| #089 | Date mentions | General | This visualisation does not show the data but a calendar. The font size of the days in the calendar depends of the frequency those dates are mentioned within the data. When a day is not mentioned within the data, it is not visible in the visualisation. | #EY063 |
| #090 | | Sensitivity | The user can choose to display only dates mentioned more than X times within the data. | #EY064 |
| #091 | Help | | Each visualisation should contain a help icon where the user can obtain instructions about the current visualisation and tips how to adjust and interpret it. We can also consider integrating this feature into the adaptive training support. | #EY032 |

| #092 | Document pane | | All visualisations should provide a document pane listing the documents according to the current search settings. The document pane should contain the (default) option to include all documents in the visualisation. The user should be able to select one, multiple or all documents. Upon selecting documents, all visualisations should update accordingly. | #EY030 |
|---|---|---|---|---|
| #093 | Filter | Geographical area | Based on the search results, a world map displays any geographic data contained by markers on the map. The user can limit the search results to specific locations by selecting and unselecting them. | #EY054 |
| #094 | Search feature | | There should be a search function to locate certain keywords within the visualisations. | #EY056 |
| #095 | Drill down | | At any stage of the analysis, the user needs to drill down to the detailed data. When selecting a concept (e.g., a tag in the tag cloud or a node in the network graph), the user should be able to receive a list of the corresponding documents, including the links to open these documents. | #EY057 |
| #096 | Export | | To facilitate the review from engagement executives, the user should be able to export the visualisations. In the export, the work done within the program should be documented, i.e., the search settings and the steps performed to adjust the visualisation. | #EY058 |
| #097 | Over time development | | The network and tag cloud visualisation also show the evolution, development | #EY042 |

| | | | and degeneration of concepts over time (taking into account the document dates) and thus allow the assessment of the relevance of a concept. | |

## 6.5 Requirement: document search and analysis

**Table 32:** Document search and analysis.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#098** | Full-text access | | Providing of full-text access. | #TUD102 |
| **#099** | Full-text search | | It should be possible for the user to search for information and keywords within the full-text of a resource he has found. | #TUD014, #TUD068 |
| **#100** | Colour of keywords | | Automatic displaying the keywords in the text with a colour. | #TUD073, #TUD100 |
| **#101** | OCR | | Text recognition on the platform. | #TUD074 |
| **#102** | Marking in the text | | Make colour marks within texts on the platform. | #TUD103 |
| **#103** | Tagging in the text | | Tagging of words, section, phrases with keywords within the text on the platform. | #TUD105 |
| **#104** | Linking of key words within one text | | Automatic linking of key words (same words) within the text and through clicking on it jumping to the next phrase or paragraph with this key word. | #TUD069 |
| **#105** | Linkage of key word within more texts | | Automatic linking of key words (same words) within texts on my search list and through clicking on it jumping to the next phrase or paragraph with this key word. | #TUD070 |
| **#106** | Comparison of documents | | Visual comparison of documents side by side. | #TUD087, #TUD088 |

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| | | | | #TUD089 #TUD111 |
| **#107** | Entity identification | (Sub-) document analysis | Entity identification in preparation for the visualisations: Extraction of entities, locations, persons and other top concepts from the data. | #EY016 |
| **#108** | | Abbreviations | The concept identification algorithm should recognise abbreviations and treat the abbreviated and the written-out word as one concept. This especially applies to laws and regulations (e.g., "HGB" and "Handelsgesetzbuch"). | #EY050 |
| **#109** | | Uniqueness | As the entity identification is key to the EY use cases, we should consider including "dictionaries" from DUNS or ISIN to take advantage of the uniqueness of those systems. | #EY060 |
| **#110** | | Excel files | As journal entry descriptions usually contain multiple words, the program should be able to separate multiple words stored in MS Excel cells. | #EY061 |
| **#111** | | Abbreviations | As there can be several abbreviations contained in the data, the tool should contain a dictionary to recognise common abbreviations. | #EY062 |

## 6.6 Requirement: video search and analysis

**Table 33:** Video search and analysis.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#112** | Annotations in videos | | Generating annotations for specified videos. | #TUD108 |

## 6.7  Requirement: Adaptive Training Support

**Table 34:** Adaptive Training Support.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#113** | Thesauri/Synonyms | | The user should be informed about which words or search terms frequently occur with each other through the adaptive training support for the MOVING platform. | #TUD003 |
| **#114** | | | The user should be informed about frequently occurring authors during the research. | #TUD010 |
| **#115** | Recommendation | | Getting recommendations of further steps while conducting the search due to the search by other users. | #TUD122 |
| **#116** | Going to discussion page | | Recommendation of going to the discussion page of the source/topic. | #TUD076 |
| **#117** | Search list | | When searching for specific paragraphs or laws and regulations, the ATS should remind the user that it might be useful to sort the search list by publication date in order to address recent changes and comments that might be relevant. | #EY052 |

## 6.8  Requirement: community

**Table 35:** Community.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#118** | Network | Contacting other researchers | It should be possible to contact other researchers via the platform. | #TUD033 |

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#119** | | Asking questions | Possibility of asking questions. | #TUD121 |
| **#120** | | | It should be possible to inform one about other researchers. | #TUD009 |
| **#121** | Pictures | | Sharing found pictures due to the search on the platform with other user of the platform. | #TUD115 |
| **#122** | Literature | | Sharing of found literature due to the search on the platform with other user of the platform. | #TUD116 |
| **#123** | Ranking | | Ranking of found resources due whether or not the result fit to the search query. | #TUD117 |
| **#124** | Discussion Forum on search results | | Possibility of discussing sources found in the search results. | #TUD076 |
| **#125** | Discussion Forum on availability of literature | | Discussing availability of literature. | #TUD127 |
| **#126** | Author details | | When searching for an author, it should be displayed whether the author has created a profile on the platform, which publications the author has published, where the author is cited and who the author cites (which is done automatically). | #TUD124 |

## 6.9 Requirement: user management

**Table 36:** User management.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#127** | Tracking queries | | Tracking of queries on search terms, search results, used databases. | #TUD078 |
| **#128** | Tagging of queries | | Tagging of queries for finding them easier. | #TUD082 |
| **#129** | Meta data library | | Meta data library for tagging search results with keywords. | #TUD079, #TUD097, #TUD099 |
| **#130** | | | Saving the search results. | #TUD095 |
| **#131** | | Reference management system on the platform | Organising and comparing the search results in a reference management on the platform. | #TUD083, #TUD096, #TUD104, #TUD077, #TUD111, #TUD087 #TUD088, #TUD089 |
| **#132** | | | Marking the found literature with colours. | #TUD106 |
| **#133** | BibTex plugin | | BibTex plugin to transfer the reference automatically, which I found on the platform into a document outside the platform. Other referencing plugins, such as RIS, Text and EndNote, could also be added. | #TUD114 |

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#134** | Documentation of search terms | | Tracking of which search term was used and in which manner (Boolean operation). | #TUD080 |
| **#135** | Notes | | Writing notes or excerpts on texts/sources. | #TUD084 |
| **#136** | | | Copy and paste these notes. | #TUD086, #TUD085 |
| **#137** | | | Download the notes. | #TUD086, #TUD085 |
| **#138** | | | Save the notes. | #TUD086, #TUD085 |
| **#139** | | Reference management system outside the platform | Exporting the search results into reference management system, which is not included on the platform. | #TUD098 |
| **#140** | Upload of resources | | Uploading resources. | #TUD110 |
| **#141** | Downloading resources | | Downloading resources. | #TUD109 |
| **#142** | Managing access of uploaded resources | | Managing access of uploaded resources into open access or not. | #TUD110 |
| **#143** | Collaborative text creation | | Creating of texts, which can be accessed collaboratively. | #TUD119, #TUD120 |

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#144** | | Save settings | Save search settings into search profiles or favourites for later use (e.g., the next year's understanding of the entity and the environment). | #EY013 |
| **#145** | Search profile management | Import/Export settings | Option to import and export search settings in order to share them with colleagues or use them for documentation purposes. | #EY014 |
| **#146** | | Predefined settings | Provide pre-defined search settings for the different uses of the platform. For example, for the ISA 315 scenario, the profile should always include the company website and a hint from the adaptive training support also to include the latest management report and notes to the financial statements available (as files from the local device). | #EY015 |

## 6.10 Requirement: accessibility of the platform

**Table 37:** Accessibility of the platform.

| ID | Topic | Key word | Description | reference to use case |
|---|---|---|---|---|
| **#147** | Accessibility to the platform from different devices | | Accessing the platform from different devices on. | #TUD107 |

## 7 Mock-ups of the interaction workflows

Sketches or mock-ups as interaction workflows in line with the key path scenarios (Noessel et al., 2014) can be used to portray how the users of the MOVING platform accomplish their goals. We started to build mock-ups of the platform in October 2016 also to collect input from the partners, to have a common visual understanding of the platform and to discuss with the project consortium. Since then, these mock-ups were discussed and modified in special online-calls in November 2016 and February 2017 and on the MOVING project meetings in November 2016 and March 2017. Next to this a wiki page in the MOVING wiki was created were all versions of the existing and former mock-ups are stored and could be discussed.

This section describes the existing mock-ups we developed with the Balsamiq Software[84]. These will be updated in D1.3. These mock-ups do not show the actual design, but serve as an orientation for the development and the implementation of the requirements. As Noessel et al. (2014) are saying the testing of sketches like this can also be useful in the very early framework phase. We plan to test these sketches in April 2017 with the help of a template for the collection of user questions and problems arising during the adoption and training of end users (see annex 9.3). Next to this, the following mock-ups present not all requirements we collected as seen in Section 6. First, the requirements need to be discussed with the developers of the platform on whether and how they can be realised and to what extend they need to be modified. Second, these sketches serve above all as an image of how the structure of the platform could look like and how all parts of the requirements and the scenarios in the DoA can be interlocked with each other.

The following figure (Figure 8) shows the general alignment of the difference functionalities with the faceted search on the left, the visualisations in the middle and the adaptive training support on the right side of the application.

---

[84] https://balsamiq.com (2017-03-27)
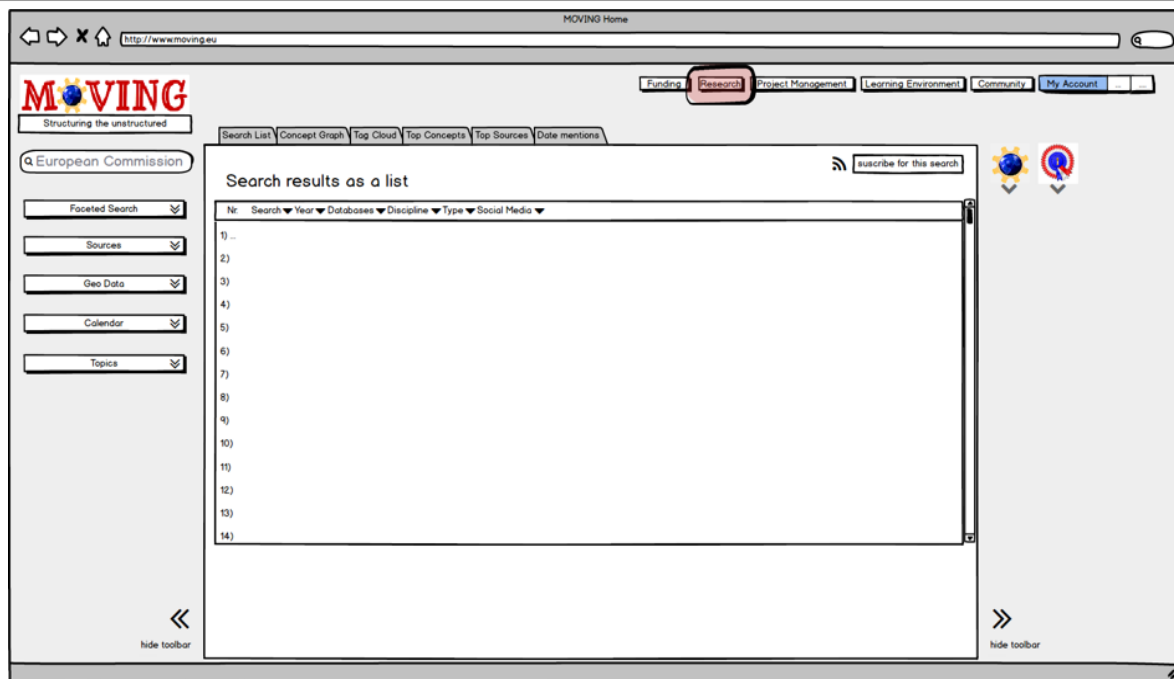
**Figure 10:** MOVING platform in "Research" view.

The upper part of the platform presents the user navigation, which aligns with the scenarios described in the DoA (funding, research, project management, learning environment and community). Next to this, the user can go to his/her profile and settings under "My account". On the left side is the search area, which contains the search input field, the faceted search, sources, Geo Data, calendar and topics (see Section 7.1 for detailed information). The mid-section of the platform presents the visualisation form of the search results. The user can switch between a conventional search list, a concept graph, a tag cloud, top concepts, top sources and date mentions (see Section 7.2 for detailed information). The right side shows the adaptive training support bar which contains the adaptive training support for the MOVING platform (left badge) and the adaptive training support for the MOVING curriculum (right badge) (see Section 7.3 and D2.1: Initial conceptual framework, curricula and technical prototypes for adaptive training support for more information on the ATS). The following subsections show the components included in more detail.

## 7.1 Faceted search
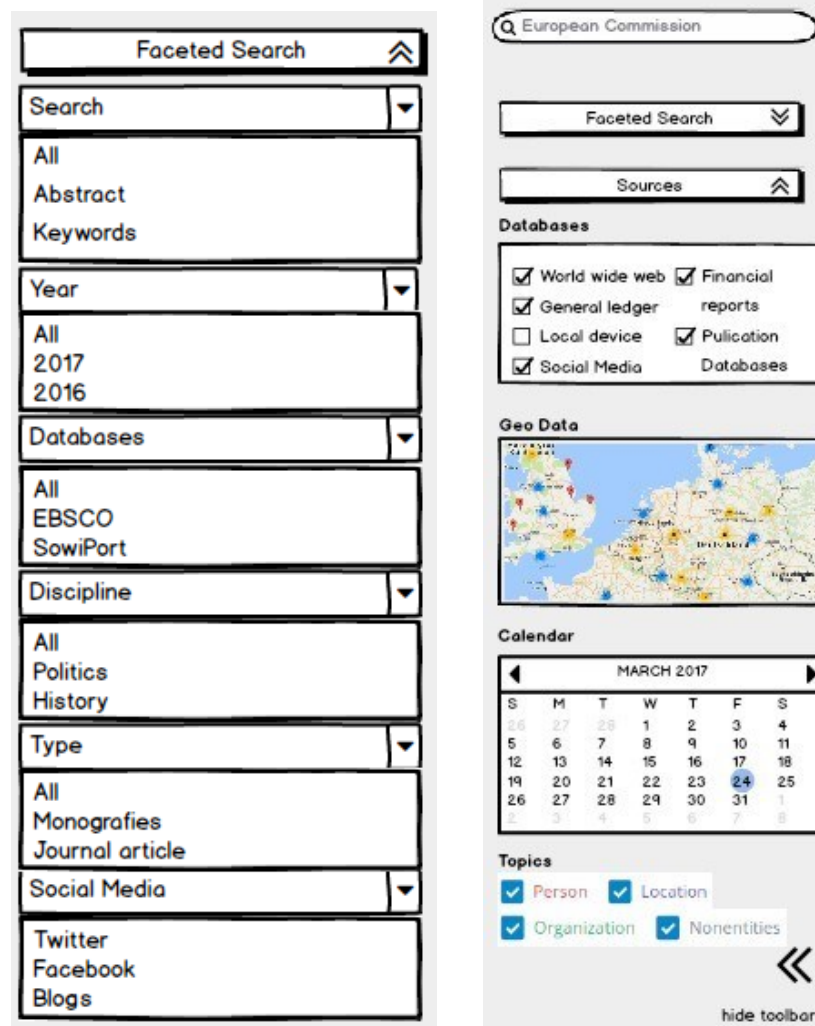


**Figure 11:** Faceted search, sources, geo data, calendar and topics.

When clicking on "Faceted search" in the left section of the platform a list of facets concerning the search is shown. This list shows possible facets at this stage of the project, which are not completed so far and not tested on usability. Another possible view instead of the drop-down list could be a checkbox view for every single facet.

Search settings can be saved into search profiles, refering to Sections 4.1, 4.2, 4.5 and 4.6 for detailed descriptions on what should be included for a more effective search experience.

## 7.2 Visualisation

The visualisations are located in the central frame of the MOVING platform. The user can switch to different visualisations by using the respective tabs. When entering a keyword in the search box or changing the search settings, the visualisations update automatically.

### 7.2.1 Concept graph
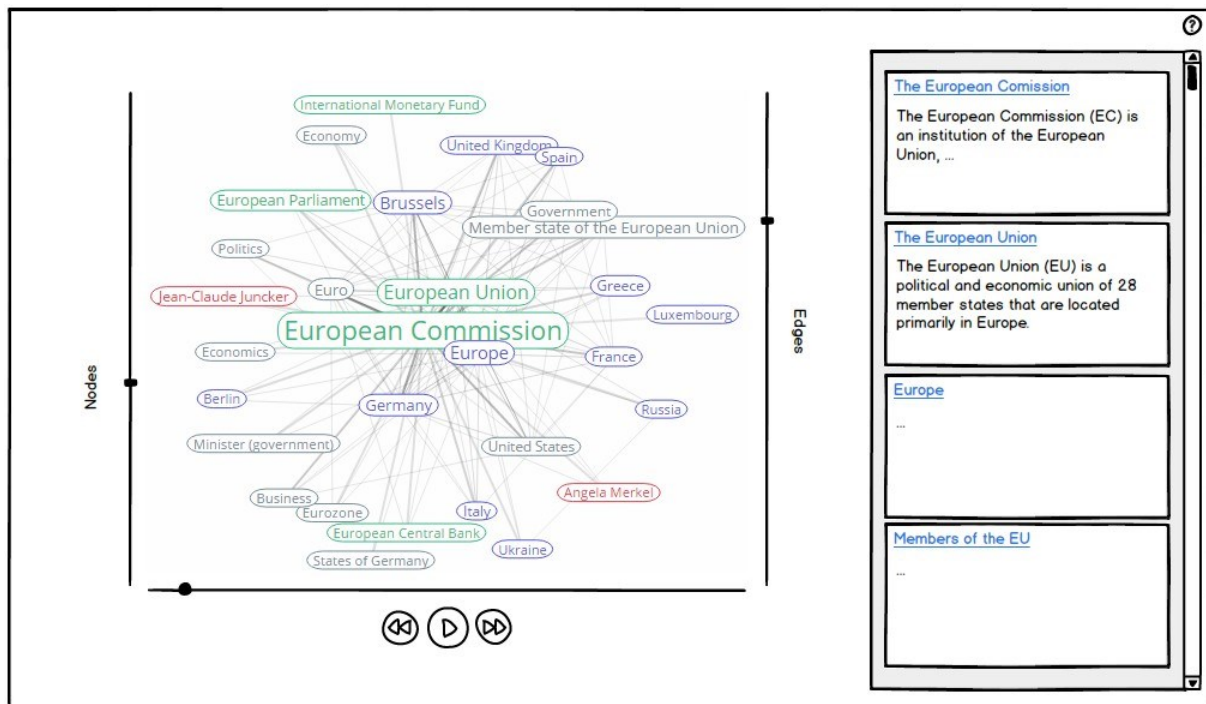


**Figure 12:** Concept graph.

The concept graph visualisation shows topics around a specific keyword. The graph consists of notes and edges linking those nodes. The nodes are the most relevant concepts within the data and the edges link the concepts (nodes) that frequently co-occur within the data. On the right side of the bar chart, the document pane is shown (details in requirements table in Section 4.1).
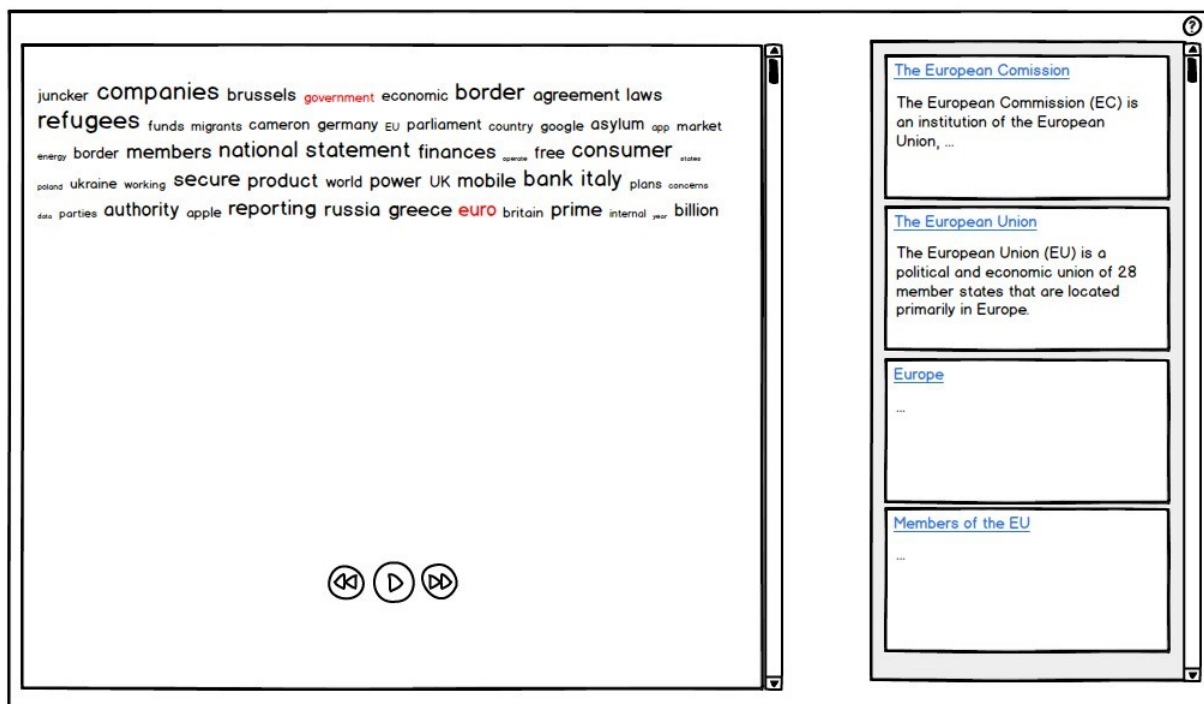
### 7.2.2 Tag cloud



**Figure 13:** Tag cloud.

The tag cloud visualisation displays the top keywords extracted from the data. The initial size of the tag depends on the tag's frequency within the search results. The initial position of the tag within the tag cloud depends on the keyword's relevance. However, there should also be an option to order the tags alphabetically. When reviewing the initial tag cloud, the user can "rate" the different tags according to his research question/area of interest. Rating tags leads to a refresh of the tag cloud taking into account these ratings. On the right side of the bar chart, the document pane is shown (details in requirements table in Section 4.1).

### 7.2.3 Top concepts
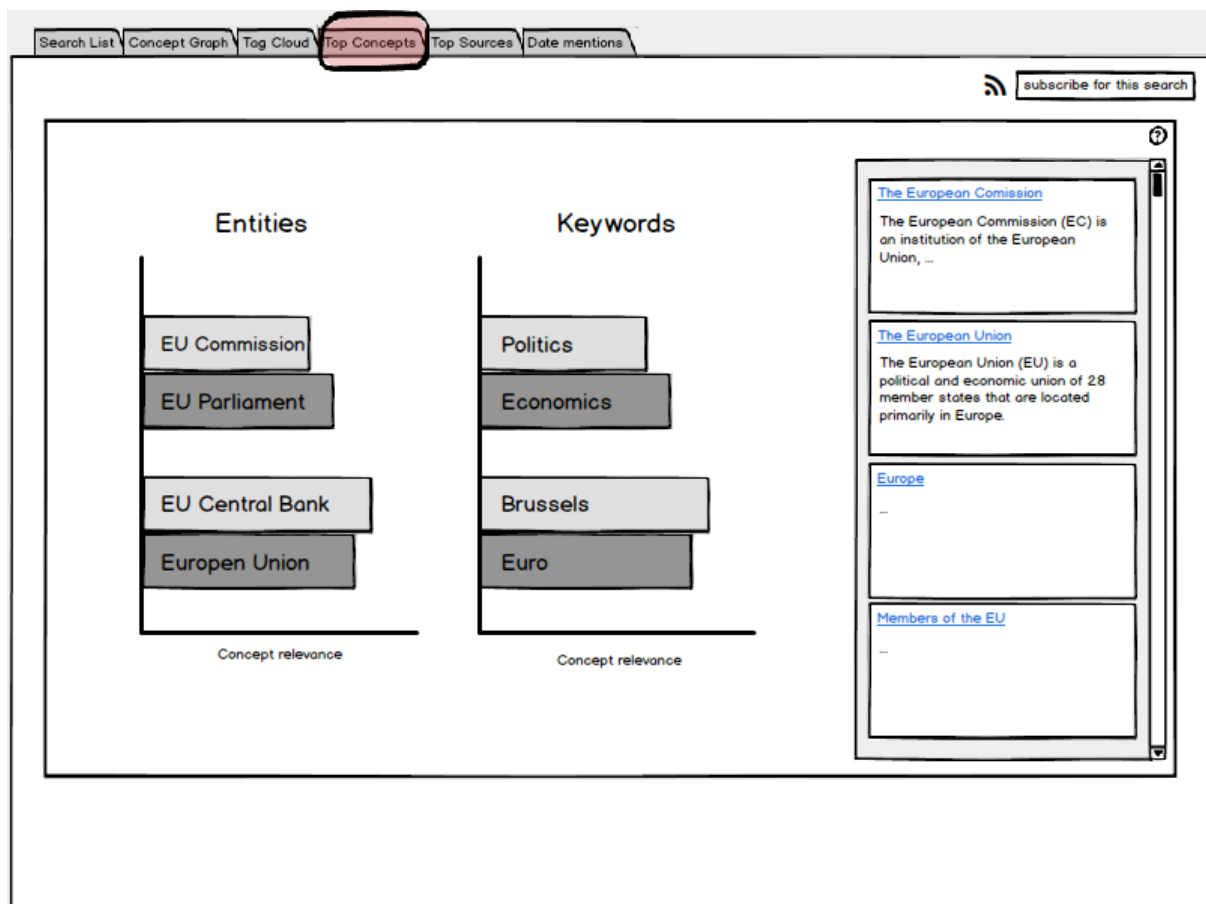


**Figure 14:** Top concept.

The top concept visualisation displays two bar charts (one for the entities, one for other concepts) that show the concepts that best describe the data. The length of each concept's bar depends on the relevance of the concept (details in requirement table in Section 4.4). On the right side of the bar charts, the document pane (details in requirement table in Section 4.1) is shown.
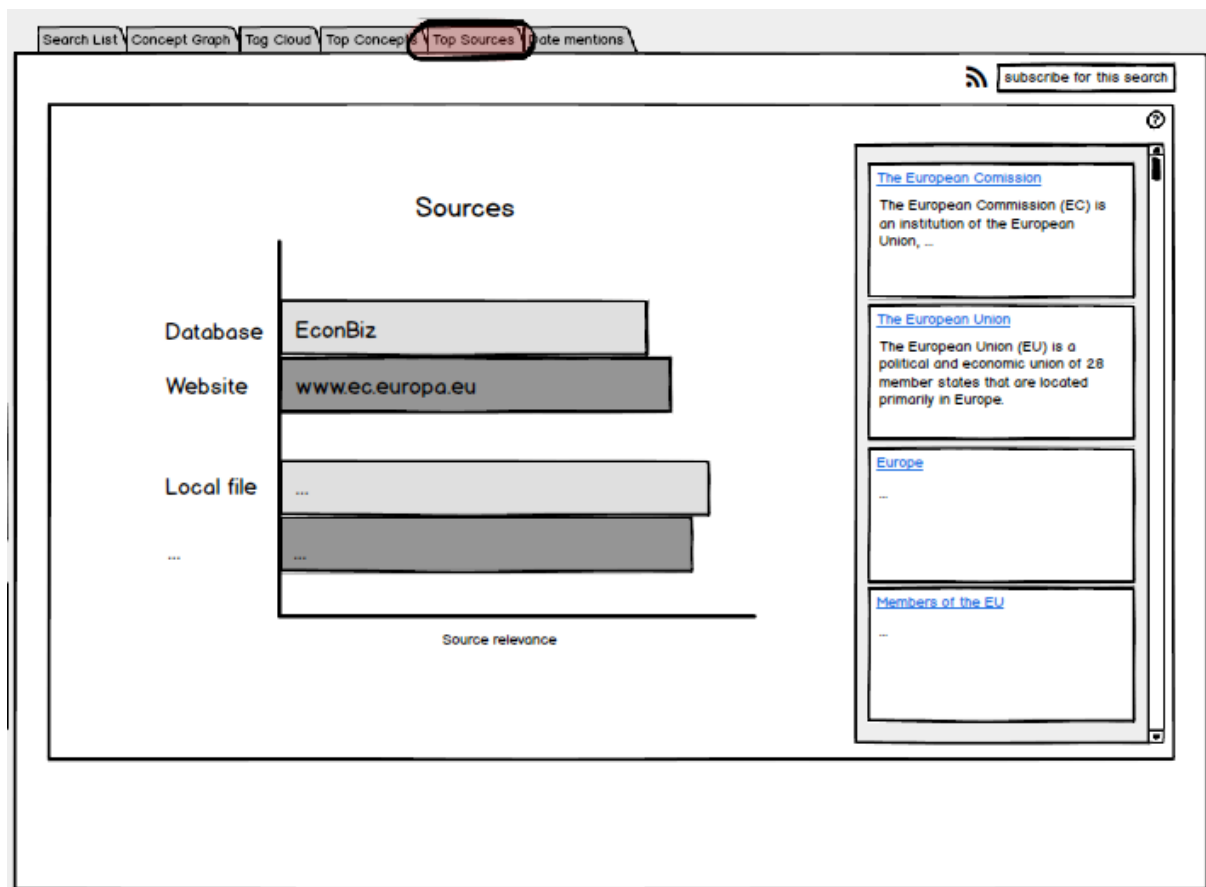
### 7.2.4 Top sources



**Figure 15:** Top sources.

The top sources visualisation shows the most relevant data sources for the given search query. The length of each concept's bar depends on the relevance of the source (details in requirements table in Section 4.4). On the right side of the bar chart, the document pane (details in requirements table in Section 4.1) is shown.
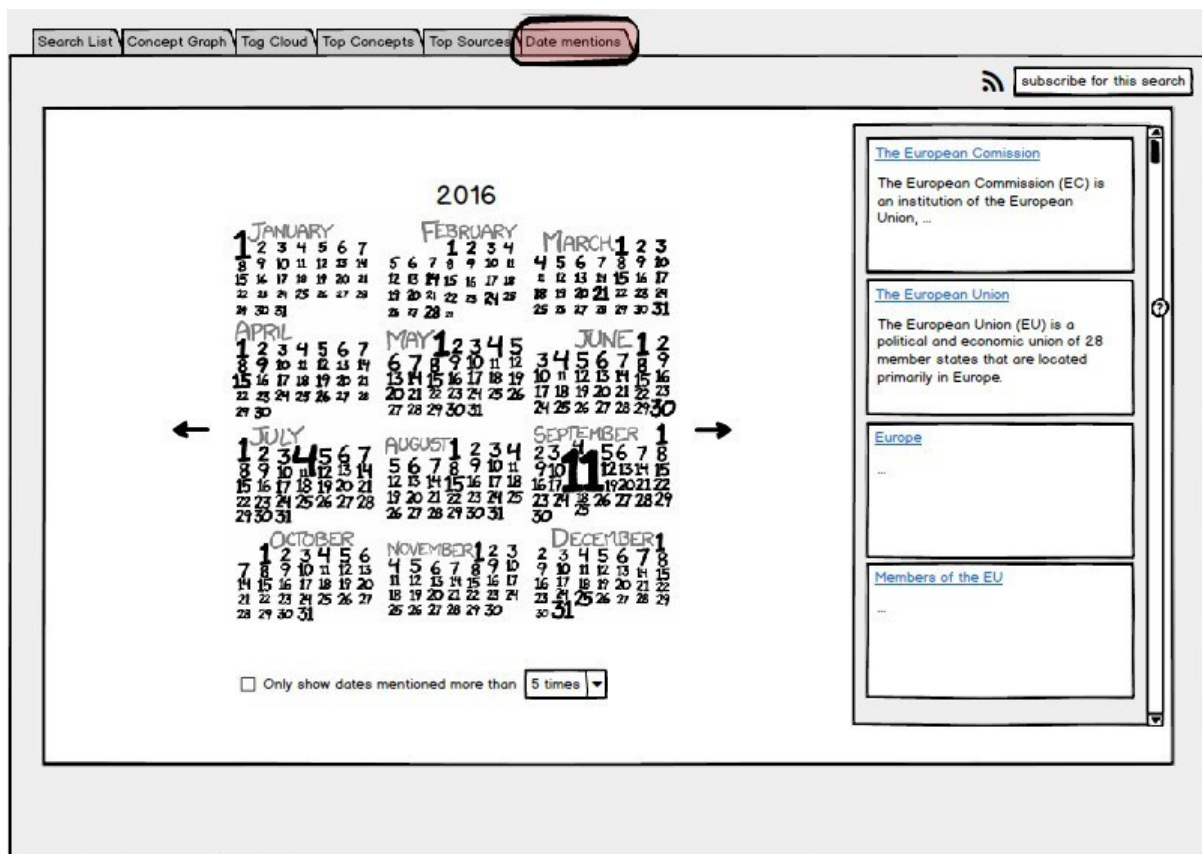
### 7.2.5   Date mentions



**Figure 16:** Date mentions.

The date mentions visualisation shows a calendar with the font sizes depending on the frequency of the date within the data set (requirements table in Section 4.7). On the right side of the bar chart, the document pane (requirements table in Section 4.1) is shown.

## 7.3   Adaptive Training Support

The ATS pursues to achieve the following two major goals: support for learning on how to use the MOVING platform and support for learning content (i.e. to become an information savvy professional). In order to reach these goals, the Adaptive Training Support consists of two types: (a) Feature-based adaptive training support to support the MOVING users how to use features available on the MOVING platform in order to improve the user's search behaviour. (b) Content-based adaptive training support to provide learning material (e.g. documents, videos) related to the user's current context and relevant for learning.

The ATS is integrated in the ATS bar on the right side of the MOVING platform, as depicted in Figure 10 and Figures 17-23, and is implemented in form of a widget. This feature-based ATS widget presents the last five used features on the MOVING platform, a performance indicator providing an opportunity for comparison as well as reflective prompts in form of questions and sentence starters.

The content-based ATS additionally recommends relevant content or topics referring to the curriculum.

A detailed description of the feature-based as well as content-based Adaptive Training Support, including its underlying concept based on literature, the provided functionality and the corresponding architecture can be found in D2.1: Initial conceptual framework, curricula and technical prototypes for adaptive training support in Section 4.

## 7.4 Funding environment
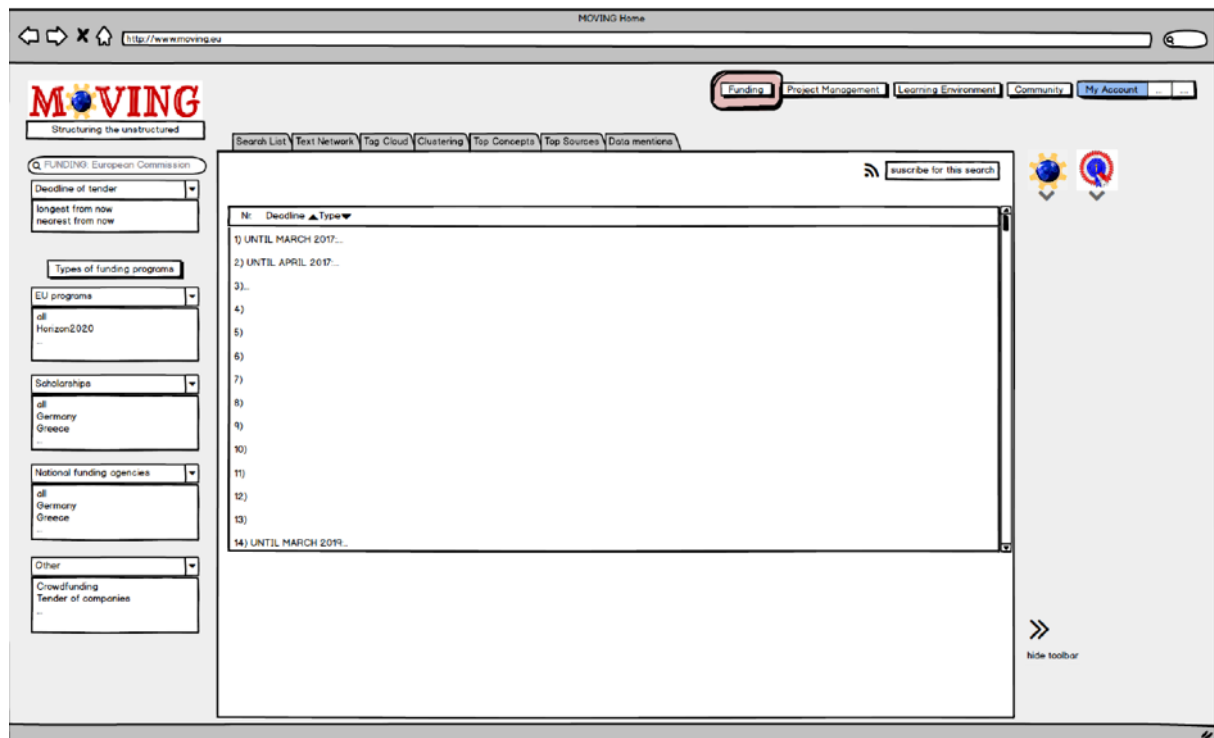


**Figure 17:** MOVING platform in "Funding" view.

This is the platform with the "Funding" functionality, which differs from the "Research" view in the left column, in the faceted search. The search bar is adjusted to the funding scenario of the second use case. Important aspects of the funding scenario include among others and as described in Section 6.3 of the requirement analysis the deadline of the specific funding option.

## 7.5 Community environment
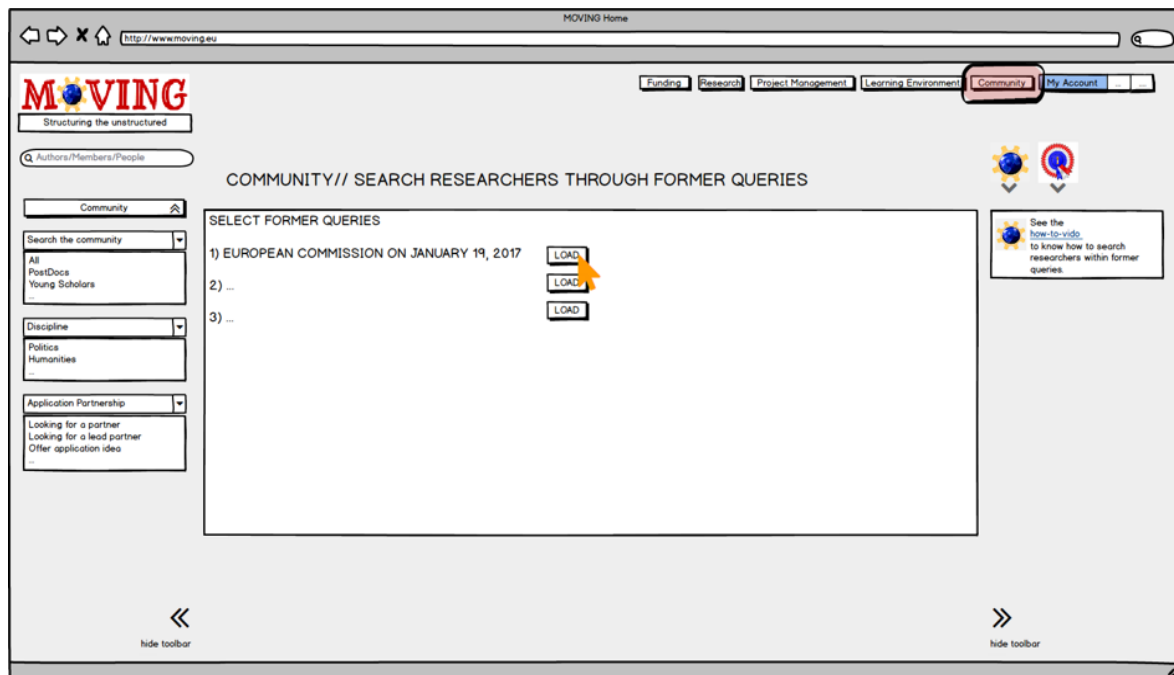


**Figure 18:** MOVING platform in "Community" view.

Figure 18 shows the "Community" functionality on the platform. By clicking on the community button in the navigation area, the user gets to this view. Through the search field, it is possible to search for authors, members on the platform and people in general. Through the search bar underneath the user can refine the search. Next to this, the user can also search authors according to the former queries he/she conducted. This is also illustrated in Figure 18. On the right side of the platform the ATS suggests a how-to-video to get to know how to search for researchers or other persons due to former queries. In this view, the user decides to search for authors not via the search field and the faceted search rather via his/her former query by clicking on "load" next to the conducted search.
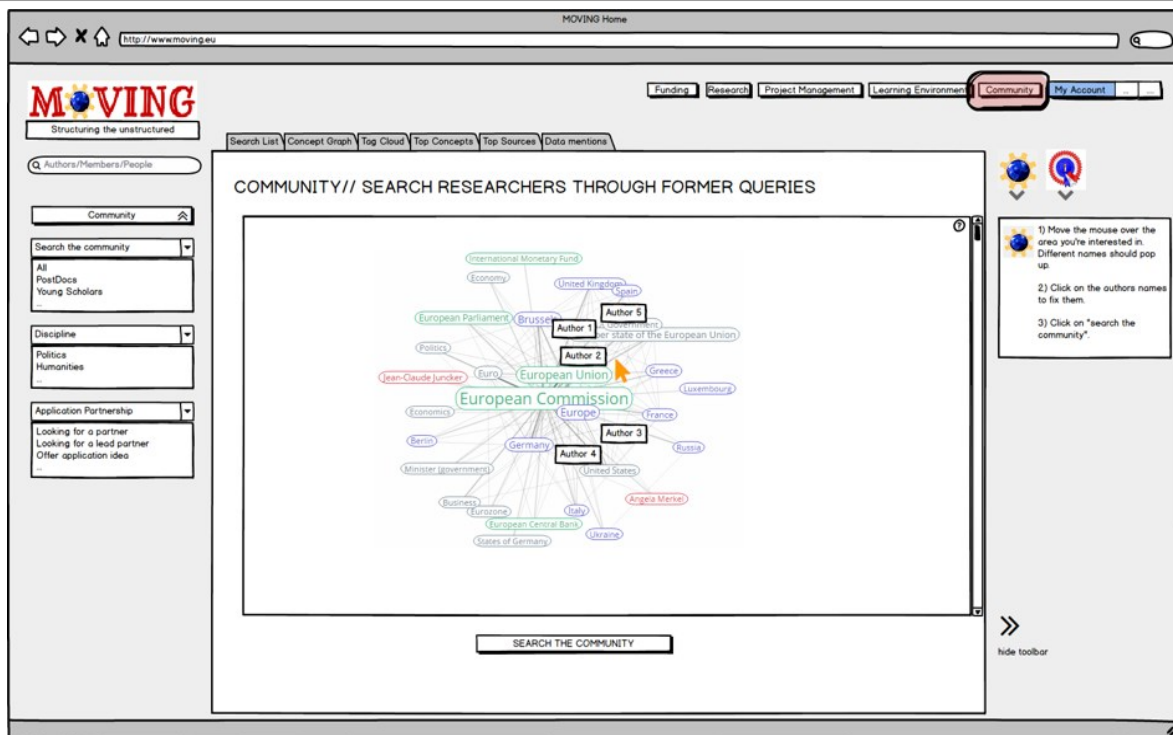
**Figure 19:** Finding authors through visualisation.

Now the user finds the concept graph he/she was looking at when conducting the research. By mouse over names of authors, pop up. The user can now fix the authors names by clicking on them. When pressing on the button "search the community" underneath the visualisation the platform will be searched in terms of the fixed author names.
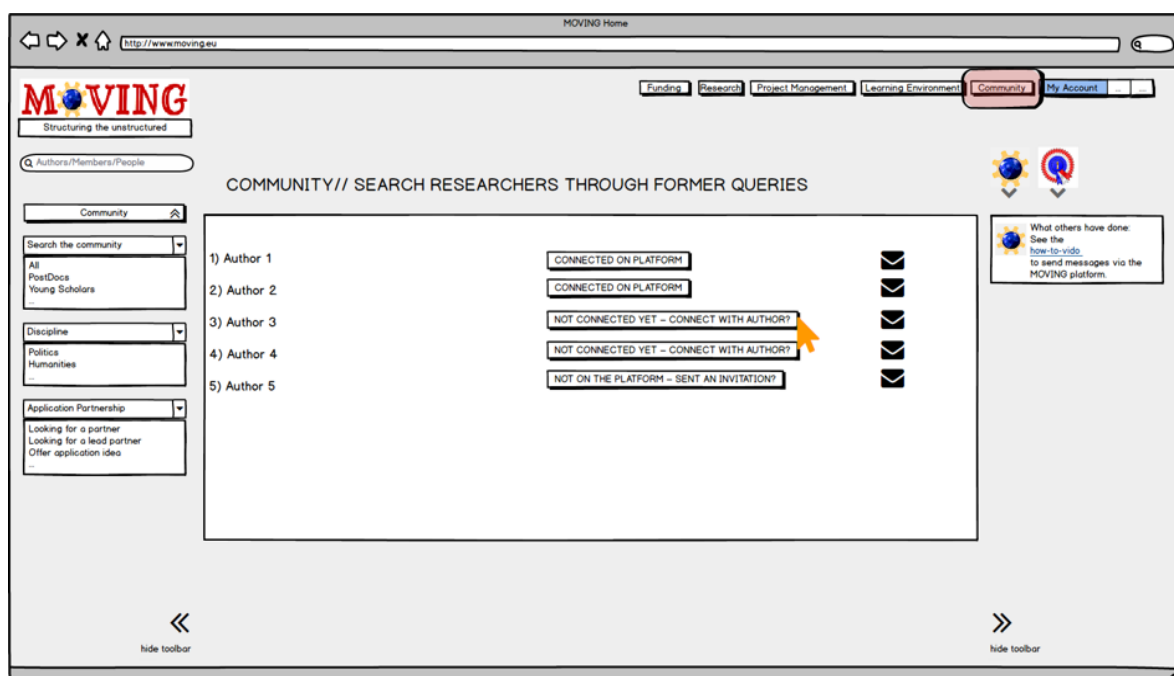
**Figure 20:** Connecting with authors.

When the search is finished, a list with the authors will be displayed in the middle of the platform. There, it is indicated whether the user is already connected to him/her, whether the user is not

connected with the author who is on the platform or whether the author is not on the platform and the user wants to invite him to join. In any case, it will be possible for the user to send a message to the author.

## 7.6 Learning environment
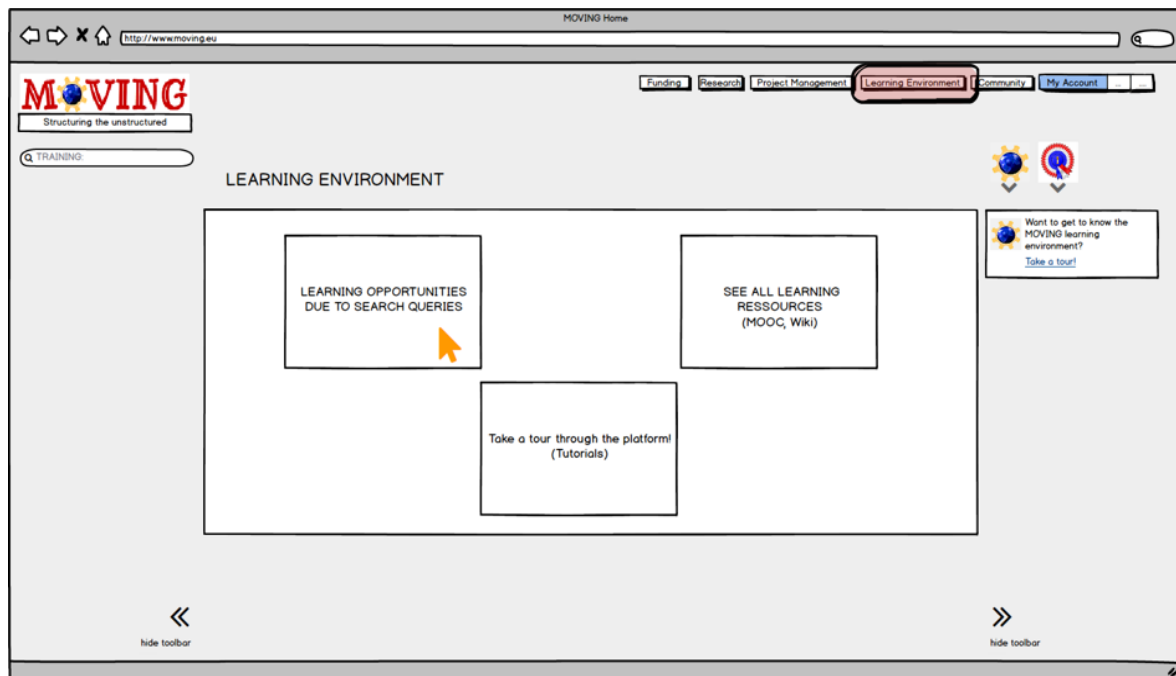


**Figure 21:** MOVING platform in "Learning environment" view.

Figure 21 shows the platform when the user clicked the button "Learning Environment" in the navigation area. Like in the "Community" functionality, the user has the opportunities to search for content through the search field and to get recommendations for learning opportunities due to former search queries. In this scenario now, the user has also the possibility to see all learning resources and to take a tour through the platform via tutorials. In this, case the user chooses to get learning opportunities due to his/her former search queries.
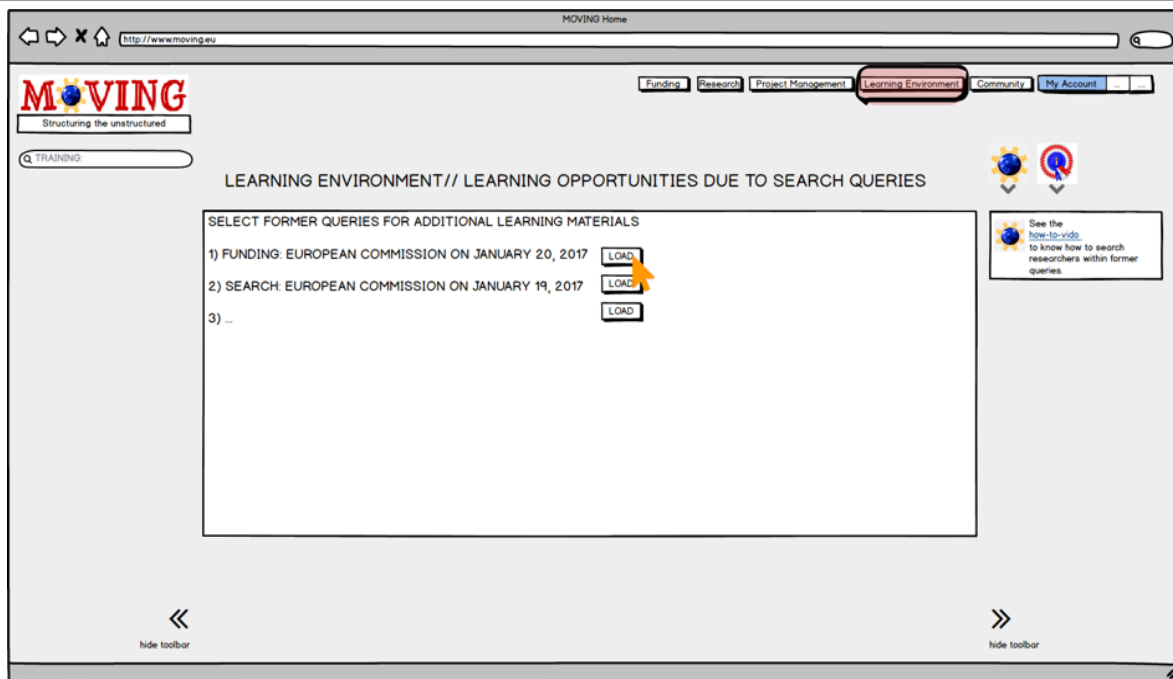
**Figure 22:** Learning opportunities due to former search queries A.

The user can now select the search query, for which he/she wants learning resources.
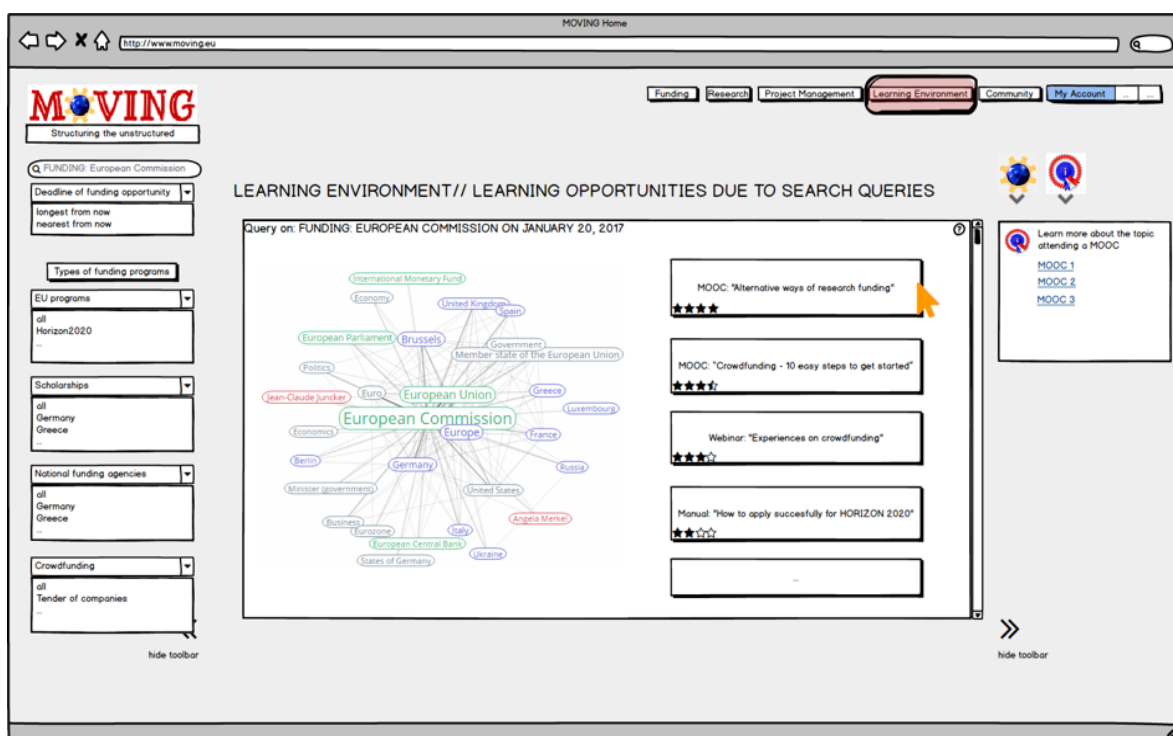


**Figure 23:** Learning opportunities due to former search queries B.

By clicking on the search query, the user gets to this search. What is different now from the regular search list/visualisation is that learning resources are displayed next to the search list/visualisation. In this case, now the user searched formerly for funding within the European Commission and specialised the search with "Crowdfunding". The search results are therefore focused on learning

resources on Crowdfunding. The stars under each resource indicate the rating of the content by other users of the platform. By clicking on the resource, the user will be guided to it.

# 8   Conclusion

This document presents the work that was done in WP 1 within the first year of the MOVING project. In this part of the deliverable, we sum up the essential initial requirements of the platform and give an outlook on ongoing tasks (Task 1.2, 1.3 and 1.4).

A mixed-method approach of collecting user requirements to the MOVING platform was chosen to identify user needs, information searching strategies and knowledge acquisition procedures. Regarding use case 1, we followed the approach of using interviews and brainstorming sessions. This led to the extensive description of the use case 1 requirements in conformity with the daily work needs of auditors and the audit process, which needs to comply with the ISA (Section 2.3). For use case 2, interviews with young scholars were conducted (Section 3.2) and translated into user stories (Section 3.2.4), which are applicable to the platform. Next to this, we reviewed related work for collecting requirements in both use cases (Section 2.4 and Section 3.3).

Based on these empirical findings, the use case requirements were developed (Section 4 and 5), which were then translated and combined in functional requirements to the platform (Section 6). These initial requirements concerning the platform are: (a) a search field, (b) a faceted search, (c) a search list, (d) visualisation options, (e) document search and analysis options, (f) video search and analysis options, (g) Adaptive Training Support, (h) community option, (i) user management option and (j) accessibility options. The interaction workflows presented in Section 7 show a possible structure in which most of the collected requirements are combined.

In the second year, the project will continue with focusing on the human-centred design approach, which was chosen for the elicitation of the requirements. With formative evaluations like testing the prototype within the end user groups and collecting user questions and problems during the adoption of training, the MOVING projects effort will focus on the Task 1.4 and continue with the implementation of both use cases as described in Task 1.2 and 1.3.

The requirements are meant to help bootstrap the first year's prototype of the platform and will be updated by D1.3 in month 24, which will contain initial evaluation, updated requirements and specifications.

# References

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 213–218). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=57203

Clarke, S. (2016). How Content and Text Analytics can be deployed to gain Competitive Advantage, Ovum TMT Intelligence.

DESTATIS Statistisches Bundesamt (2015). Bildung und Kultur. Studierende an Hochschulen - Fächersystemtatik. Retrieved from: https://www.destatis.de/DE/Methoden/Klassifikationen/BildungKultur/StudentenPruefungsstatistik.pdf?__blob=publicationFile

Döring, N. & Bortz, J. (2016). Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Berlin/Heidelberg: Springer

Evelson, B. (2015). "Vendor Landscape: Big Data Text Analytics", Forrester. Retrieved from: http://goascribe.com/wp-content/uploads/2015/05/Vendor_Landscape__Big_Dat.pdf

Franke, F. (2016). Standards der Informationskompetenz - neue Entwicklungen in Deutschland, Großbritannien und den USA. In: Sühl-Strohmenger, W. (Ed.): Handbuch Informationskompetenz, 2nd Ed., pp. 22-29. Berlin: De Gruyter.

Gupta, V. & Lehal, G.S. (2009). "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence Vol. 1, No. 1, August 2009. Retrieved from http://learnpunjabi.org/pdf/gslehal-pap18.pdf

Hargreaves, H., Guibault, L., Handke, C., Seville, P. & Seville, B. (2014). "Standardization in the area of innovation and technological development, notably in the field of Text and Data Mining", European Commission Report. Retrieved from https://www.iprhelpdesk.eu/printpdf/2459

Klingenberg, A. (2016a). Referenzrahmen Informationskompetenz. Retrieved from: http://www.bibliotheksverband.de/fileadmin/user_upload/Kommissionen/Kom_Infokompetenz/2016_11_neu_Referenzrahmen-Informationskompetenz_endg__2__Kbg.pdf.

Klingenberg, A. (2016b). Referenzrahmen Informationskompetenz für alle Bildungsebenen. In Sühl-Strohmenger, W. (Ed.): Handbuch Informationskompetenz, 2nd Ed., pp. 30-41., Berlin: De Gruyter.

Kramer, B. & Bosman, J. (2016). Innovations in scholarly communication - global survey on research tool usage [version 1; referees: 2 approved] F1000Research 2016, 5:692. doi: 10.12688/f1000research.8414.1

Kuckartz, U. (2010). Einführung in die computergestützte Analyse qualitativer Daten. Wiesbaden: VS Verlag.

Mayring, P. (2010). Qualitative Inhaltsanalyse. Grundlagen und Techniken, Weinheim: Beltz.

Noessel, C.N., Cronin, D.C., Reimann, R.R., & Cooper, A.C. (2014). About Face: The Essentials of Interaction Design, 4th Edition. John Wiley & Sons.

Parenteau, J., Sallam, R.L., Howson, C., Tapadinhas, J., Schlegel, K. & Oestreich, T.W. (2016). "Magic Quadrant for Business Intelligence and Analytics Platforms", Gartner. Retrieved from http://www.webmining.cl/wp-content/uploads/2015/02/mqbi2015.pdf

Tristram, F., Bamberger P. & al. (2016): Öffentlicher Abschlussbericht von bwFDM-Communities. Wissenschaftliches Datenmanagement an den Universitäten Baden-Württembergs, Online available: http://bwfdm.scc.kit.edu/downloads/Abschlussbericht.pdf

Tullis, T. S., & Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. In *Usability professional association conference* (pp. 1–12). Retrieved from https://pdfs.semanticscholar.org/9d62/1a71a9c9e24f689f5263ed5b74e53535374a.pdf

Wirdemann, R. (2011). Scrum mit User Stories, München: Hanser Verlag

Yuhanna, N. (2015). "Market Overview: Graph Databases", Forrester.  Retrieved                from: https://www.forrester.com/report/Market+Overview+Graph+Databases/-/E-RES121473

# 9 Appendix

## 9.1 Interview guide for the requirement analysis of use case 2: "Managing and mining research information"

**Part I Approaches, proceeding, strategies and experiences with the use of unstructured data and digital/web-based tools**

1. The increasing importance of the internet as an information resource is accompanied by an ever-growing availability of data and information (e.g. publications, research data, information from websites and social media such as blogs, twitter or social networks). What does this mean for your research work and how important does this become for your discipline, from your point of view?
2. How do you proceed with the ever-growing amount of data and information?
   a. What do you expect thereby from searching the internet?
   b. Why do you use especially the internet or web-based tools in order to inform yourself?
3. When and in what working process is searching after information important to you?
4. What kind of digital/online available information and data (e.g. texts, videos, pictures etc.) are important within your working area or your discipline?
   a. Where do you find them? From what resources do you get them?
5. Now we will focus on specific working processes from your daily research routine, where searching and processing information is important. Please describe your working processes in detail. How do you proceed the processes below in order to search information on the internet and utilise them for your research work?
   a. Imagine you are approaching a new research topic: How would typical steps look like to get first information on that topic? Please describe them.
      i. Where and how do you search after information?
      ii. When you have found information, how do you assess/verify that this information is relevant to your research topic?
      iii. When you have found an interesting information, how do you further proceed it?
   b. Imagine after you approached a new research topic, you are now searching for research funding opportunities (e.g. to conduct a collaborative research project on that topic or to fund your personal research work). How would typical steps look like, to find information on funding opportunities?
      i. Where and how do you search after information?
      ii. How do you proceed to assess/verify relevant information?
      iii. When you have found an interesting information, how do you proceed with it?
   c. If you want to work on a newly approach research topic collaboratively with other researchers (e.g. within a research project), how would typical steps look like to find them?
      i. Where and how do you search for the specific information?
      ii. How do you proceed to assess/verify which researcher you want to collaborate with or it is expedient to collaborate with?

      iii.   When you have found a suitable researcher, how do you further proceed?

    d.  How important is the search after the following kinds of information within your daily research routine and how do you proceed to find information on them:

      i.   Regulatory framework of research work (e.g. labour law, copyright, data security).

      ii.   Conferences, other scientific events, fairs, summer schools.

      iii.   Stay abroad for research.

      iv.   Scientific exchange (e.g. networks).

      v.   Specific tools for research (e.g. clouds, reference management).

6. Do you use specific methods or proceedings in order to search and proceed information?

    a.  Searching methods/strategies such as:

      i.   Snowball method.

      ii.   Network analysis.

      iii.   Linguistic methods (e.g. collocation analysis).

      iv.   Text Mining.

      v.   "Follower Power" (i.e. to ask for other users' help, e.g. on twitter, using the hashtag #followerpower).

    b.  Methods to depict or structure information:

      i.   Mind-Mapping or other methods of information visualisation.

      ii.   Tagging.

7. What digital/web-based tools or websites do you use, to search after information and proceed them?

    a.  What tools do you like or do you favourably use?

    b.  If you have a favourite tool, describe why you prefer it.

    c.  Where did you find these tools?

    d.  What devices do you use?

8. What kind of challenges are you faced with in your research routine regarding unstructured data (e.g. data volume, structure, resources, access), where do you perceive opportunities for improvement and needs for support?

    a.  Where do you observe opportunities for improvement regarding tools?

    b.  Which working processes should be connected or should work automatised?

      i.   E.g. automatised visualisation of searching results.

    c.  How such a scenario could look like?

9. If a web-based tool/platform for searching and processing research information would be developed, are there functionalities that you regard as necessary (e.g. network visualisation, tag clouds, filter functions, thesauri)?

10. What is the general framework and what are constraints of your work with large amounts of unstructured data (e.g. copyright, privacy, and data security)?

## Part II Questions on information literacy and the usage of web-based learning and training offers

The second part of the interview focuses on training offers, to improve information literacy of researchers. Thereby information literacy is understood as to involve the following competencies: (a) Searching (i.e. express a need for knowledge, find suitable resources and choose and isolate information), (b) assessing/verifying regarding thematic relevance, factual correctness, formal

accuracy and completeness, (c) generating knowledge (i.e. express, compare, arrange and structure information), (d) describing/depicting regarding understandability, (e) sharing information (i.e. clarify terms of use, indicate citations, name resources, use networks).

1. What kind of learning/training offers did you already attend, to improve one of the named competencies (e.g. during your study or as further training; as online training or face-to-face teaching)?
2. In general terms: How do you train yourself? What training formats do you use and which do you prefer?
3. What experiences do you have with digital training formats? Which of the following training formats do you know which do you already have attended and how satisfied have you been: Blended Learning, Game Based Learning, MOOCs, self-learning modules, web-based training.
4. What else digital training formats do you know and have experiences with?
5. How useful are digital training formats for you in general?

## 9.2 Scheme of categories from the qualitative data analysis

**Main categories**

As main categories, the following four scenarios as defined in MOVING were used:

| | | |
|---|---|---|
| **Scenario 1** | State-of-the-art on a research topic | |
| **Scenario 2** | Finding suitable partners for research projects | Finding suitable partners for research projects that are active in the respective research field. |
| **Scenario 3** | Finding suitable research funding | Strategic decision for deciding to go for which research funding for my topic. |
| **Scenario 4** | Training | Accompanying training materials, courses, tutorials. |

Additional scenarios have been defined, if a user story could not clearly be categorised as part of one of the scenarios pre-defined in the DoA:

| | | |
|---|---|---|
| **Scenario 5** | Collaborative work | Collaborative work on a research project together with different research partners. |
| **Scenario 6** | Networks | Building up a personal network or use networks. |

**First level sub-categories**

Within every main category, (scenario) sub-categories have been pre-defined. On the first level, pre-defined sub-categories are used according to six aspects of information literacy, as defined within the German framework of reference for information literacy. These are the following:

| | |
|---|---|
| Searching | Searching for information on a research topic, this involves:<br><br>• Expressing the need for further knowledge in that topic.<br><br>• Finding resources.<br><br>• Choosing resources.<br><br>• Segregating information. |
| Assessing/verifying | Assessing/Verifying information regarding the relevance for the research topic/the own research work, this involves:<br><br>• Relevance of the topic.<br><br>• Factual accuracy.<br><br>• Formal accuracy.<br><br>• Completeness. |
| Generating knowledge | Generating knowledge out of information through:<br><br>• Formulating/phrasing (i.e. to re-express a relevant information in technical language or with own words).<br><br>• Comparing (i.e. to express own assumptions or answers on a research question based on new information from different resources; to compare new and existing information or put them into a broader context).<br><br>• Integrating/re-arranging (i.e. to develop own opinions or reconsider them, based on new information; to integrate information into the overall context.<br><br>• Structuring. |
| Delineating | Delineating information through:<br><br>• Linguistic simpleness (appropriately illustrate a topic for the target group).<br><br>• Semantic redundancy (to perceive fundamental propositions and to put them in a meaningful context or transfer them into another context). |

| | |
|---|---|
| | • Cognitive structuring (to parse a topic distinctively).<br><br>• Cognitive conflict (to arouse interest on a topic). |
| Sharing | Sharing information with others, this involves to:<br><br>• Clarify terms of use.<br><br>• Mark citations.<br><br>• Name resources.<br><br>• Use networks for sharing. |

Regarding scenario four (training), the following sub-categories were defined on the first level:

| | |
|---|---|
| Teaching and learning methods, techniques or formats | With which teaching and learning methods, techniques or formats are the interviewees familiar or have experiences with? |
| Self-regulated learning | How do the interviewees adopt knowledge out of formal learning contexts? Which resources do they use? |
| Training in the field of information literacy | What kind of courses on information literacy (or aspects of it) have the interviewees already attended, e.g. courses as part of the curriculum of their study or training courses offered by the university or the library, online or face-to-face courses? |

**Second level sub-categories**

On the second level, sub-categories were defined related to the different aspects of information literacy (level 1):

| Sub-category level 1 | Sub-category level 2 | Description |
|---|---|---|
| Searching | Tools for searching | Tools used for searching. |
| | Search content | Kind of the content or the unstructured data the interviewees searched for (e.g. text, videos). |
| | Organising search results | To organise search results or information |
| | Searching strategies | Description of how the interviewees proceeded their searching (systematic descriptions or description of single steps). |

| Assessing/ verifying | Assessing/ verifying strategies | Description of how the interviewees proceeded assessing/verifying the relevance and/or reliability of an information (systematic descriptions or description of single steps). |
|---|---|---|
| | Reliability | Assessing/verifying the reliability and credibility of information and resources. |
| | Relevance regarding research topic | Assessing/verifying the relevance of information or resources regarding the research topic. |
| Generating knowledge | Text-mining | Text mining as a method of analysing information to generate knowledge. |
| Delineating | Mind-mapping | Delineating information with the help of mind-maps as specific kind of visualisation |
| | Visualisation | Delineating information with the help of visualisations. |
| Sharing | Ethics | Ethical questions of scientific research and how they are considered regarding the sharing of information and knowledge. |
| | Copyrights/ Terms of use | Consider copyright and terms of use of information and content in the process of sharing. |
| | Personal rights | Consider personal rights in the process of sharing. |

Regarding the scenario "finding suitable research funding" (scenario 3) it has been further refined for what kind of funding researchers are searching:

| Sub-category level 1 | Sub-category level 2 | Description |
|---|---|---|
| Searching | Foundations | Search for foundations to find suitable funding. |
| | Scholarships | Search for scholarships as funding opportunity. |

Regarding the scenario "training" (scenario 4) different teaching and learning techniques, methods or formats have been further refined:

| Sub-category level 1 | Sub-category level 2 |
|---|---|
| Teaching and learning methods, techniques or formats | Videos. |
| | Game-based-Learning. |

| | | |
|---|---|---|
| | Face-to-face teaching (e.g. workshops or courses). | |
| | MOOCs. | |
| | Flipped Classroom. | |
| | Webinar. | |
| | Blended learning. | |

**Third level sub-categories**

Some second level sub-categories had to be further refined with the help of a third sub-category:

| Sub-category level 2 | Sub-category level 3 | Description |
|---|---|---|
| Tools for searching | Social networks | Social networks are used for searching information. |
| | Search engine | Search engines are used for searching information. |
| | Databases | Specific scientific databases are used for searching for information. |
| | Online archives | Specific scientific online archives are used for searching for information. |
| | Journals | Search in specific scientific journals. |
| | Websites | Search on websites (to use the "search" function of websites). |
| | Newsletters | Subscribe to newsletters to find information. |
| | Video portal | Use video portals (e.g. YouTube) for searching. |
| Search content | Online encyclopaedia | Search for content in online encyclopaedias (e.g. the Wikipedia). |
| | Text | Search for text. |
| | Videos | Search for videos. |
| | Audio | Search for audio content (e.g. podcasts). |
| Organising | Reference management | Use any kind of reference management for |

| search results | | organising search results. |
|---|---|---|
| | Institutional server | Use institutional servers for organising search results. |
| | Tagging | Tag search results with keywords. |
| Searching strategies | Snowball principle | Search for literature or persons based on cited authors within a publication, while reading further publications, further persons and literature can be opened up. |

## 9.3 Template for the collection of user questions and problems arising during the adoption and training of end users

- Overall impression and reaction to the platform:
- How would you describe the aim of the platform in 20 words?

|  |
|---|
|  |

- Please indicate on the following scales your reaction to the platform (Chin, Diehl & Norman 1988):

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
|---|---|---|---|---|---|---|---|---|---|---|
| terrible | | | | | | | | | | wonderful |
| difficult | | | | | | | | | | easy |
| frustrating | | | | | | | | | | satisfying |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **inadequate power** | | | | | | | | | | **adequate power** |
| **dull** | | | | | | | | | | **stimulating** |
| **rigid** | | | | | | | | | | **flexible** |

- Interaction with the platform
- Please indicate on the following 5-step scale to which extend you agree on the following statements (adapted from Tullis & Stetson, 2004).

| | Strongly disagree | Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|
| **I think that I would like to use this platform frequently.** | | | | | |
| **I found the platform unnecessarily complex.** | | | | | |
| **I thought the platform was easy to use.** | | | | | |
| **I think that I would need the support of a technical person to be able to use this platform.** | | | | | |
| **I found the various functions in this platform were well integrated.** | | | | | |
| **I thought there was too much inconsistency in this platform.** | | | | | |
| **I would imagine that most people would learn to use this platform very quickly.** | | | | | |
| **I found the platform very cumbersome to use.** | | | | | |
| **I felt very confident using the platform.** | | | | | |
| **I needed to learn many things before I** | | | | | |

| could get going with this platform. | | | | | |
|---|---|---|---|---|---|

- Had you any problems while interacting with the platform?
    a. Yes
    b. No
- If yes, which problems arose while you interacted with the platform?

| |
|---|
| Title of problem 1:<br><br>Section of the site (see title of the browser):<br><br>In detail description of the problem:<br><br><br><br> |
| Title of problem 2:<br><br>Section of the site (see title of the browser):<br><br>In detail description of the problem:<br><br><br><br> |
| Title of problem 3:<br><br>Section of the site (see title of the browser):<br><br>In detail description of the problem:<br><br><br><br> |
| Title of problem 4:<br><br>Section of the site (see title of the browser):<br><br>In detail description of the problem:<br><br><br><br> |
| Title of problem 5:<br><br>Section of the site (see title of the browser):<br><br>In detail description of the problem:<br><br> |

Title of problem 6:

Section of the site (see title of the browser):

In detail description of the problem:

## 9.4 Databases

| Fachportal Pädagogik | http://www.fachportal-paedagogik.de/metasuche/erweiterte_suche.html#zusaetzlich |
| --- | --- |
| Springerlink | http://link.springer.com |
| Archive.org | https://archive.org |
| GESIS database catalogue | https://dbk.gesis.org/dbksearch/home.asp?db=e |
| DBOD database | https://www.dbod.de/ |
| LEXISNEXIS | http://e-solution.lexisnexis.de/KSH/de/index.html |
| BDSL | http://www.bdsl-online.de/BDSL-DB/templates/template.xml?vid=%7B92382D71-12C8-4994-A8C0-F5915E474B9B%7D&contenttype=text/html&Skript=home&lang=de |
| Germanistik | http://www.germanistik-im-netz.de/dbis/ |
| WILEY | http://onlinelibrary.wiley.com |
| EBSCOHost | https://www.ebscohost.com |
| DigiZeitschriften | https://www.digizeitschriften.de/startseite/ |
| JSTOR | https://www.jstor.org |
| DBIS (Datenbank-informationssystem) | http://dbis.uni-regensburg.de//fachliste.php?lett=l |
| Bundesarchiv | http://www.bundesarchiv.de/rech erche/index.html.de |
| Deutsche Fotothek | http://www.deutschefotothek.de |

## 9.5   Library catalogues

| SLUB | https://www.slub-dresden.de/startseite/ |
|---|---|
| German National Library | http://www.dnb.de/EN/Home/home_node.html |
| Library of Congress | https://www.loc.gov |

## 9.6   Websites

| Qualitative research | http://www.qualitative-research.net/index.php/fqs |
|---|---|
| | forum function on the page "Forum qualitative social research" |
| OECD | http://www.oecd.org |
| WTO | https://www.wto.org |
| Perseus | http://www.perseus.tufts.edu/hopper/ |
| H-Soz-Kult | http://www.hsozkult.de |
| Gesetze im Internet | https://www.gesetze-im-internet.de |
| Flickr | www.flickr.com |