2016-09-19

Using Spatialisation to Support Exploratory Search Behaviour

Clement Roux
Technological University Dublin

Follow this and additional works at: https://arrow.dit.ie/scschcomdis

Part of the Computer Engineering Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the School of Computing at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@dit.ie, arrow.admin@dit.ie, brian.widdis@dit.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.
Using Spatialisation to Support Exploratory Search Behaviour

Clément Roux

A dissertation submitted in partial fulfilment of the requirements of

Dublin Institute of Technology for the degree of

M.Sc. in Computing (Data Analytics)

August 2016
DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute’s guidelines for ethics in research.

Signed: ____________________________

Date: 31 August 2016
ABSTRACT

Information-seekers traditionally interact with digital content through keyword-based search interfaces displaying results in list views. Well-defined lookup search tasks are performed brilliantly with these interfaces, enabling users to find relevant information and develop a relative understanding of the underlying information space. However, it is feasible to suggest that ill-defined and abstract search tasks could be better supported with a different interface that could allow the user to explore a library’s content and develop an appropriate mental model of the information space. One such approach is based on the use of visualisation, an approach to data analysis that aims to reduce cognitive burned by capitalising on perceptual capabilities. One common approach to visualising a large collection of documents is based upon a spatialisation which translates high dimensional spaces into 2D planes, where each item’s location reflects its relationships with the rest of the library’s content.

This research seeks to establish whether a spatialisation of digital libraries’ content can influence users’ exploratory search behaviour. To do this, a between-group online experiment was conducted to measure respondents’ levels of sensemaking accuracy, exploratory search behaviour and cognitive load while interacting with a novel spatialisation interface, called ExploViz and its non-visual equivalent, called LibSearch. Results show that the respondents exhibited similar levels of exploratory search behaviour irrespective of interface. However, a slightly significant improvement, supporting the project’ hypothesis, was observed while performing the sensemaking task using the ExploViz interface. These results posit interesting questions about how and at what stage exploratory search tasks could be more effectively supported with more visualisation-based interfaces.
ACKNOWLEDGMENTS

I would like to express my sincere gratitude and thanks to my supervisor Dr. John McAuley for his continuous support, motivation, invaluable feedback, patience and knowledge.

Also, thanks to Dr. Eric Briys, founder of Cyberlibris, for sharing his data, his ideas, interest, knowledge and time for the project.

Finally, the kindest thanks goes to my family for their support and to Flore great and supportive as always.
CONTENT
Declaration............................................................................................................... I
Abstract......................................................................................................................II
Acknowledgments ..................................................................................................... III
Content...................................................................................................................... IV
List of figures .............................................................................................................. IX
List of tables .............................................................................................................. XI
1 Introduction ........................................................................................................... 1
  1.1 Background ...................................................................................................... 1
  1.2 Research project and problem ...................................................................... 1
  1.3 Research objectives ....................................................................................... 2
  1.4 Research methodology ............................................................................... 3
  1.5 Scope and challenges ................................................................................. 3
  1.6 Contributions .............................................................................................. 4
  1.7 Document outline ....................................................................................... 4
2 Background .......................................................................................................... 6
  2.1 Background domains .................................................................................... 6
    2.1.1 Libraries ............................................................................................... 6
    2.1.2 Information retrieval ............................................................................ 7
      2.1.2.1 The Boolean and probabilistic ranking methods ....................... 7
      2.1.2.2 The semantic models .................................................................. 8
    2.1.3 Information seeking: Lookup search and exploratory search .......... 8
      2.1.3.1 Lookup Search .......................................................................... 8
      2.1.3.2 Exploratory search .................................................................... 9
    2.1.4 Information visualisation ...................................................................... 13
      2.1.4.1 The definition ............................................................................ 13
      2.1.4.2 The design studies .................................................................... 14
    2.1.5 Data analytics and exploration ............................................................ 14
    2.1.6 Data visualisation and exploration ....................................................... 15
2.1.7 The role of visualisation in digital repositories and libraries .......... 16

2.2 Design considerations ................................................................................. 19
  2.2.1 The information space ........................................................................... 19
  2.2.2 Spatialisation of metadata ..................................................................... 20
  2.2.3 Topic modelling ...................................................................................... 26
  2.2.4 Nice, easy and intuitive to use ................................................................. 31
  2.2.5 “Overview first, zoom and filter, then details-on-demand” ...................... 36
  2.2.6 Interactivity ............................................................................................ 39

2.3 Summary ........................................................................................................ 40

3 Design and implementation ............................................................................. 42
  3.1 The ExploViz interface design: a visualisation design study .................... 42
    3.1.1 Domain problem characterisation ........................................................ 43
    3.1.2 Data and task abstraction design – What and Why? ......................... 45
      3.1.2.1 The tasks – Why? ............................................................................ 45
      3.1.2.2 The data – What? ............................................................................ 46
    3.1.3 Encoding and interaction technique design – How? ............................. 49
      3.1.3.1 Spatialisation ................................................................................... 50
      3.1.3.2 Marks and channels ....................................................................... 50
      3.1.3.3 Interactivity .................................................................................... 51
        3.1.3.3.1 Manipulate – select and navigate .............................................. 51
        3.1.3.3.2 Reduce - embed ........................................................................ 52
  3.2 The ExploViz interface implementation ...................................................... 54
    3.2.1 Data collection ...................................................................................... 55
    3.2.2 Data profiling and sample selection from the original dataset .......... 55
    3.2.3 Data cleansing and data standardisation .............................................. 55
      3.2.3.1 Data import in R ........................................................................... 56
      3.2.3.2 Text cleansing .............................................................................. 56
      3.2.3.3 Dataset semantic enrichment ......................................................... 56
3.2.4 Data transformation – Data spatialisation .................................................. 58

3.2.4.1 The document term matrix .......................................................................... 58

3.2.4.2 The Latent Dirichlet Allocation (LDA) topic modelling. .......................... 59

3.2.5 The ExploViz graphical user interface ......................................................... 61

3.3 Summary ........................................................................................................... 62

4 Evaluation ........................................................................................................... 63

4.1 The experimental design .................................................................................... 63

4.1.1 A between group experiment comparing two search interfaces ................. 64

4.1.2 An online task-based experiment .................................................................. 64

4.1.3 The user-tasks ............................................................................................... 65

4.1.3.1 Sensemaking task 1 .................................................................................. 66

4.1.3.2 Directive scenario-based task 2 ............................................................... 66

4.1.3.3 Free exploratory search task 3 .................................................................. 67

4.1.4 The cognitive load measurement .................................................................. 67

4.1.5 The evaluation metrics .................................................................................. 68

4.1.5.1 Measuring exploratory behaviour ............................................................. 68

4.1.5.1.1 The number of hover events ............................................................... 69

4.1.5.1.2 The Area explored .............................................................................. 69

4.1.5.2 Syntactic and semantic proxies for sensemaking accuracy .................... 70

4.1.5.2.1 Syntactic sensemaking proxy ............................................................... 70

4.1.5.2.2 Semantic sensemaking proxy ............................................................... 71

4.1.5.3 The distance between two selected books .............................................. 72

4.1.5.4 The exploration time .................................................................................. 72

4.2 The baseline search interface: LibSearch ......................................................... 73

4.3 The experiment implementation ....................................................................... 75

4.4 Results .............................................................................................................. 76

4.4.1 Data preparation ........................................................................................... 76

4.4.2 Data manipulation ......................................................................................... 77

4.4.3 Statistics ........................................................................................................ 78
7.1.3 Cognitive load measurement across ......................................................... 106
7.1.3.1 The respondents’ germane cognitive load ........................................ 106
7.1.3.2 The respondents’ extraneous cognitive load ..................................... 107
7.2 Resulting visual spaces of Document-Term Matrix initial transformations.... 108
LIST OF FIGURES

Figure 2.1 The “search activities” .................................................................................. 9
Figure 2.2 The iterative search and the exploratory search ............................................ 10
Figure 2.3 Exploratory search diagram ........................................................................ 13
Figure 2.4 The four-level nested model ........................................................................ 14
Figure 2.5 The mid-level action of a user-task when interacting with a visualisation . 16
Figure 2.6 The Universal Digital Library’s website ...................................................... 16
Figure 2.7 The Spinger’s AuthorMapper: ...................................................................... 17
Figure 2.8 The 3D Vase Museum .................................................................................. 21
Figure 2.9 The ActiveGraph ......................................................................................... 21
Figure 2.10 The Hierarchical Spatialisation Algorithm .................................................. 22
Figure 2.11 PATHS exploration tab .............................................................................. 23
Figure 2.12 The reference map ..................................................................................... 24
Figure 2.13 The starSPIRE .......................................................................................... 25
Figure 2.14 The Refinery .............................................................................................. 26
Figure 2.15 The Grisham word-cloud tab ..................................................................... 27
Figure 2.16 The Serendip ............................................................................................. 28
Figure 2.17 The LDAvis ............................................................................................... 29
Figure 2.18 The Mimir’s co-occurring term matrix ....................................................... 30
Figure 2.19 The Bookfish digital library for children ..................................................... 31
Figure 2.20 The SIZL ................................................................................................... 32
Figure 2.21 The SciNet .................................................................................................. 33
Figure 2.22 The Adaptive VIBE .................................................................................. 34
Figure 2.23 The Bohemian Bookshelf .......................................................................... 35
Figure 2.24 The WikiGalaxy home mode ..................................................................... 36
Figure 2.25 The WikiGalaxy fly mode .......................................................................... 38
Figure 2.26 The Overview ............................................................................................ 38
Figure 2.27 Jigsaw document cluster feature ................................................................. 39
Figure 2.28 Jigsaw similarity feature ............................................................................ 40
Figure 3.1 Visualisation Framework and Organisation Overview ..................................... 43
Figure 3.2 The Digital Content Explorer (DICE) implemented by Cyberlibris ........... 44
Figure 3.3 Bar chart representing the initial dataset collected from Cyberlibris .......... 47
Figure 3.4 Transformation of the dicetheme ................................................................. 48
Figure 3.5 The ExploViz visual search interface in un-zoomed mode......................... 49
Figure 3.6 The ExploViz visual search interface in zoomed mode............................... 50
Figure 3.7 Illustration of the semantic zoom configured in the ExploViz..................... 52
Figure 3.8 Illustration of the ExploViz keyword search focus + context tool.............. 53
Figure 3.9 Illustration of the ExploViz visual search interface.................................. 54
Figure 3.10 Data quality chart displaying the entry dataset for LDA ......................... 58
Figure 3.11 Sparse document-term matrix’s .............................................................. 58
Figure 3.12 Hierarchical clustering on the dissimilarity correlation matrix .................. 60
Figure 4.1 The user-tasks detailed through the multi-level task typology .................... 65
Figure 4.2 Illustration of the area explored metrics.................................................. 69
Figure 4.3 Illustration of two possible area measure.................................................. 70
Figure 4.4 Illustration of the syntactic index ........................................................... 71
Figure 4.5 The baseline list-view search interface home page................................... 73
Figure 4.6 One book’s detail page as displayed in the list-view search interface.......... 74
Figure 4.7 The first task visual instructions .............................................................. 75
Figure 4.8 The first task respondent’ keyword input and “Begin Task Two” button. ... 76
Figure 4.9 Syntactic sensemaking accuracy data manipulation for the topic 1............. 78
Figure 4.10 Syntactic sensemaking across both interfaces......................................... 102
Figure 4.11 Semantic LSA sensemaking across both interfaces .................................. 103
Figure 4.12 Semantic LDA sensemaking across both interfaces ............................... 103
Figure 4.13 Semantic Dandelion sensemaking across both interfaces ....................... 104
Figure 4.14 Number of hover events across tasks and interfaces................................ 104
Figure 4.15 Area explored measured as the surface covered by each topic’s polygon 105
Figure 4.16 Euclidean distance between two selected books.................................... 105
Figure 4.17 Exploration time of task 3 extracted in seconds..................................... 106
Figure 4.18 Germane cognitive load across tasks and interface types ....................... 107
Figure 4.19 Extraneous cognitive load across tasks and interface types ................... 108
Figure 7.1 w3_dtm + cosine_distance + 4Walktrap_community ................................ 109
Figure 7.12 w1_dtm + euclidean_distance + 5Clusters_Kmeans............................... 110
Figure 7.13 w0_dtm + jaccard_index + 5Clusters_Kmeans......................................... 111
Figure 7.14 w2_dtm + cosine_distance + 5Clusters_Kmeans..................................... 112
LIST OF TABLES

Table 2.1 Overview of the design considerations.......................................................... 41
Table 3.1 Original fields collected from Cyberlibris....................................................... 47
Table 3.2 Derived fields from the original fields. ........................................................... 48
Table 3.3 Level 1 distributions among the nine topics.................................................... 60
Table 4.1 Task 1 statistics - behavioural metrics and cognitive loads............................ 79
Table 4.2 Task 1 statistics - the four proxies measuring the sensemaking accuracy...... 80
Table 4.3 Task 2 statistics - behavioural metrics and cognitive loads............................ 82
Table 4.4 Task 3 statistics - behavioural metrics and cognitive loads............................ 83
Table 4.5 Summary of all tested p-values ...................................................................... 85
1 INTRODUCTION

1.1 Background

Libraries can be either physical or shape. The two mainstream classifications, namely the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) literally shape most of the physical libraries around the world. On the contrary, the digital libraries are not subject to the books’ classification constraint, providing them more flexibility. The significant advances in information retrieval have had great influence in the digital libraries’ development. It has particularly improved the indexing techniques and retrieval of document based on users’ requests formulated as queries. Today, both physical and digital libraries are very different and yet provide complementary ways to access knowledge. One of the main difference lies in the users’ strategies to find documents. Digital libraries are known to optimise the lookup search strategy, whereas the physical libraries probably outperform the digital libraries regarding exploratory search strategies. This project’s point of departure focuses on the latter.

Information Retrieval (IR) has had a significant impact on how we interact with digital libraries. Lookup search behaviour has been improved dramatically, enabling users to quickly find books of interest related to subject, content analysis and citation analysis. However, exploratory search behaviour has received much less attention, potentially because users have become accustomed to traditional IR interfaces, adapting their search strategies to the lookup search tools. Information visualisation, however, has shown promise in supporting and improving cognition when searching information in libraries and, given this potential, should more effectively support exploratory search behaviour (Gerken et al., 2009; White, Kules, Drucker, & others, 2006; Zaphiris, Gill, Ma, Wilson, & Petrie, 2004). Despite this potential, there is a lack of studies in the literature detailing how visualisation can support exploratory search in digital libraries. Studies that address this problem, are mainly qualitative and difficult to compare.

1.2 Research project and problem

Today’s search interfaces are mainly optimised toward precise search and question answering. However, there is a commonly accepted lack of support for exploratory search tasks which are part of information-seekers daily activities (Jon Pearce et al., 2011, p. 253; Dana McKay, Shukla, Hunt, & Cunningham, 2004, p. 283; White, Kules,
et al., 2006, p. 37). This research project addresses the visualisation aspect of the information-seeking lack of support toward exploratory search detected in search interfaces covering digital collections of documents. More specifically, the project seeks to define whether visual spatialisation impacts the exploratory search behaviour of users interacting with a digital library in an exploratory search context. An exploratory search is generally defined as a complex problem whose goal is not clearly defined and where the information-seeker is unfamiliar with the informational context (Ryen W. White & Resa A. Roth, 2009, p. 10).

Can visual spatialisation influence exploratory behaviour when compared with traditional search interfaces?

1.3 Research objectives

The project aims at contributing to explaining the role of visualisation within an exploratory search process in digital libraries. More details about the high-level and low-level objective are presented below.

The principal research objective of this project consists in augmenting human capabilities when entering an exploratory search within a digital library by using information visualisation as a support tool for improving the users’ sensemaking and exploratory search behaviour. As outlined by Munzner & Maguire (2014), visualisation “is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods” (p. 1). Indeed, today’s digital collections of documents are usually optimised for specific search tasks rather than open-ended, ill-defined, persistent and complex search tasks. Therefore, the research objective consists in testing whether information visualisation can be used as a support for augmenting human capabilities when entering an exploratory search within a digital library.

The secondary research objective consists in defining the impact of spatialisation on exploratory search tasks performed in digital libraries. Understanding the role of visualisation, and more particularly spatialisation, can be useful for improving search interfaces and assist the users in engaging in serendipitous discovery within digital collection of documents. It can be part of a general improvement of future search interfaces supporting hybrid search behaviours and strategies. Additionally, the project
aims at providing metrics for measuring users’ exploratory behaviour as well as a novel visual search interface, spatializing a collection of digital documents.

1.4 Research methodology
The research begins with a literature review, considered as a secondary piece of research. Indeed, the research gap, the research question and the visual search interface design are derived from it and leads to the project’s primary piece of research. The adopted approach is inductive and empirical as the project’s conclusions stem from experimental observations of users’ behaviour interacting with a search interface. The data underlying the primary research are collected from an online between-group task-based experiment designed so that respondents express an exploratory behaviour while performing the three search tasks. Additionally, the respondents answer Likert scale cognitive load questions. The collected data are quantitative and can be considered both objective and subjective. Indeed, the users’ behaviour captured through weblogs such as the hover events are objective data, whereas the users’ perceptions Likert scale questions are subjective. The between-group experiment is used for comparing two independent groups of respondents exposed to two distinct search interfaces.

The project is therefore divided between one human-computer-interaction piece of research and one visualisation design study. Indeed, the research question focuses on one specific search behaviour defined in information-seeking theory which leads to measuring levels of users’ exploratory behaviour, sensemaking accuracy and cognitive load. In addition, the research question also requires the implementation of a visual search interface for measuring the impact of spatialisation. The project’s methodology is therefore multifaceted as it tries to meet several requirements from distinct research domains. The design study follows the four-level nested model providing a framework and practical guidelines (Meyer, Sedlmair, & Munzner, 2012; Tamara Munzner, 2009). Additionally, the experiment follows the traditional methodology of an online between-group task-based experiment.

1.5 Scope and challenges
The project’s scope is mainly defined by its goal, which is to test empirically the impact of spatialisation on the readers’ exploratory search within a digital collection of documents. The project is shaped and scoped toward information visualisation, information-seeking and digital libraries’ metadata. Indeed, it uses one precise
information visualisation technique, namely the two dimensional spatialisation, which limits the visualisation design to a smaller design space. It also focuses on a precise information-seeking task, namely the exploratory search task, which is well-defined and opposed to the traditional lookup search. Finally, the underlying information space relates to the specific digital libraries’ metadata. Typical librarian metadata consists in the documents’ titles, authors, abstracts, covers, tables of contents and themes. Those are the only information that the project uses for implementing both search interfaces of the project. Indeed, the documents’ contents are not accessible through the search interfaces.

Testing the impact of visualisation in such explorative context is challenging because it requires to design an experiment which can capture comparative levels of exploratory search behaviours between visual and non-visual search interfaces. Those requirements therefore lead to the implementation of two search interfaces; one is a novel visual spatialisation and the other one is a traditional library search interface where spatialisation is replaced with a keyword-based search bar and a list-view. Additionally, the project contributes to designing exploratory-search user-tasks and corresponding metrics for measuring the resulting respondents’ behaviour. This is another challenge which requires controlling for users’ exploratory behaviours and designing for novel and tailor-made metrics capturing sensemaking accuracy, levels of exploratory behaviours and cognitive loads.

1.6 Contributions

The principal contribution of this research project is an empirical study to assess the impact of a digital library’s visual spatialisation on users’ exploratory search behaviour, compared to a traditional list-based search interface. This approach is new and aims at fulfilling a gap from the literature. A set of design considerations as well as a novel visual search interface are also part of the contribution.

Additionally, the research project aims at contributing a set of metrics for evaluating sensemaking accuracy and exploratory search behaviour.

1.7 Document outline

The document is divided into 5 chapters. Following this introductory chapter, the chapter 2 - Background first introduces the reader with the context of the research. It then exposes the design considerations derived from the literature and informing the
project’s visual interface design. Then, the **chapter 3 - Design and implementation** first focuses on the project’s visual interface design study derived from the design considerations previously listed and following the *four-level nested model*. It then details the interface implementation starting from the data collection and transformation into R, to the graphical user interface. The **chapter 4 - Evaluation** first exposes the experimental design, followed by details about the baseline state-of-the-art interface. It then exposes and interprets the results collected from the online between-group task-based experiment. Finally, the **chapter 5 - Conclusion** provides conclusions, limitations and a reflection about the research project.
2 BACKGROUND

The background chapter addresses both a literature review about the project’s underlying domains as well as design considerations informing the ExploViz interface design. The latter is a spatialisation, which main goal consists in turning “high-dimensional data into visualizations via processes of projection and transformation” (Skupin & Fabrikant, 2003). It uses the underlying space as a map where similar books are positioned close to each other. Additionally the spatialisation follows closely Shneiderman's (1996) mantra encouraging visualisations to display the underlying information space’s overview and to offer some details and filter options on demand.

The background chapter first discusses in section 2.1 the overall context of the research by introducing the background domains as well as several specific domains’ associations. Both are derived and reviewed from the literature. The project’s context mainly relates to libraries, information retrieval, information seeking and information visualisation, which form the four first sections of the chapter. It is followed by three specific associations, namely data analytics and exploration, data visualisation and exploration followed by the role of visualisation in digital repositories and libraries. Then, the section 2.2, discusses the visualisation design considerations reflected from the literature and partly responsible for the ExploViz interface design. The six design considerations are introduced interchangeably one after the other. Finally, the section 2.3 concludes the chapter.

2.1 Background domains

2.1.1 Libraries

Libraries are probably the biggest source of knowledge ever built by humans. Either physical or digital they are an important source of information for researchers, teachers, students and casual readers. Even though both physical and digital libraries have the same function, the ways and means of reaching information are completely different. Indeed, a physical library provides a Euclidean space of documents represented by fixed and parallel bookshelves, whereas a digital library generally provides a query-search interface, resulting in ranked lists of documents matching the information-seeker’s keywords.

The book selection process is different in physical and digital libraries especially when it comes to exploring the information space. Indeed, it was shown that navigation and
serendipity are facilitated in physical libraries even though digital libraries have the great advantage of ordering and re-ordering their bookshelves as desired (Hinze, McKay, Vanderschantz, Timpany, & Cunningham, 2012; Dana McKay & Conyers, 2010). Moreover, evidence for co-location browsing behaviour was found in physical libraries, showing that near-location items from chosen books have more chance to be picked as well. Those typical behaviours found in physical libraries are probably the consequences of its physical nature, which by definition shows the readers with bookshelves. However, those beneficial consequences are also the source of inconveniences which make digital libraries successful. Indeed, the digital libraries show well-known major advantages due to its digital nature and supported by the major advances in information retrieval. This explains why both forms of libraries are complementary, have their own characteristics, and coexist independently to some extent (D. McKay, Smith, & Chang, 2014).

2.1.2 Information retrieval
Digital libraries and information retrieval share a long history due to the major importance of the information retrieval process when searching a digital library. Along the years, the natural language processing methods used for extracting information evolved from Boolean and probabilistic ranking methods to semantic modelling techniques.

2.1.2.1 The Boolean and probabilistic ranking methods
In the late 50’s, the Boolean indexing method was the first information retrieval (IR) breakthrough. Cleverdon (1959), as cited in Sanderson & Croft (2012) demonstrated that searching a document using the keyword indexing method was more efficient than the physical libraries’ traditional hierarchical classifications (Library of Congress or Dewey Decimal System). Following the indexing method, the probabilistic ranked retrieval was discovered and its original principle still remains in use today – generally described as term frequency weighting. Each keyword attached to the documents is assigned a weight that will define the document’s relevancy against the user’s query. The weight is first defined manually but quickly the term’s occurrence frequency is used as an indicator of the keyword’s significance in a document. Since then, the retrieval performances and effectiveness have been significantly improved, using additional techniques and more complex algorithms. Later, the term frequency (TF) was associated to the inverse document frequency (IDF), forming the TF-IDF weighting technique. With TF-IDF, the more popular a term in the corpus, the more weakened the term frequency, decreasing
the final weight. Nevertheless, regardless of the specific technique, information retrieval remains an end-to-end textual process in which documents are analysed as bags-of-words and the results of the keyword-based query-search is presented as a list of documents.

2.1.2.2 The semantic models
Nowadays textual corpuses are transformed into document-term matrices allowing the corpus’ space to be perceived as vectors of terms and documents. This approach enables more semantic-focused techniques as opposed to the purely textual and syntactic methods. For example, latent semantic indexing (LSI), a popular approach developed during the early nineties, makes use of the singular value decomposition for dimensionality reduction purposes. This technique allows digital documents to be semantically positioned on a two dimensional plane forming meaningful clusters (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

More recently, Latent Dirichlet Allocation (LDA) topic modelling technique provides a probabilistic topical representation of the documents contained in a document-term matrix. LDA generates a specific number of topics, given as input parameter to the model. Each topic is made of a set of words and each document is associated with the probabilities corresponding to each topic. The model assumes that the most likely co-occurring combinations of terms have a similar meaning and constitutes a topic. One of the model’s strength is its ability to allocate each document to a probability for each topic underlying the collection reflecting the general intuition that a document can relate to multiple different topics (Blei, 2012; Blei, Ng, & Jordan, 2003).

2.1.3 Information seeking: Lookup search and exploratory search
Information seeking and information retrieval show some similarities, but mainly differentiate about the integration of the user behaviour into the theoretical models. Indeed, information seeking integrated uncertainty, cognition and specific search strategies as a component of exploratory behaviour. The following section focuses on two distinct and often complementary search behaviours that were defined in one influential paper from Gary Marchionini (2006).

2.1.3.1 Lookup Search
The lookup search can be considered as a task and as a behaviour. Task-wise, a lookup search can be seen as an information need that contains a clearly-defined goal and that
can be answered using a simple and well-formulated query. As outlined in Figure 2.1, typical lookup search activities are “fact retrieval”, “questions answering” or “verification”. The lookup task generally triggers a lookup behaviour which can be described as simple mental activities, well-defined queries and well-structured search strategies with no need for special investigation nor examination of the results. Indeed, the information-seeker has a good mental representation of the search domain being knowledgeable about the context and knowing the right keywords forming the right query (Gary Marchionini, 2006).

As explained by Marchionini (2006), [...] “lookup tasks are suited to analytical search strategies that begin with carefully specified queries and yield precise results with minimal need for result set examination and item comparison” (p. 42).

Typically, the information retrieval tools are developed specifically for optimising the lookup search activities, which are usually evaluated using precision, recall and other derivative metrics. The latter assume that each result is performed against a query and that it can be classified as true or false result.

![Figure 2.1 The “search activities” differentiate the lookup search from the exploratory search (Gary Marchionini, 2006, p. 42)](image)

2.1.3.2 Exploratory search

Beyond the analytical tasks, for which search engines are optimised, Marchionini (2006) outlines additional learning and investigative search activities that he defines as exploratory search. Learning activity can be seen as a motivation to assemble new
knowledge, which usually asks for multiple iterations, concentration and interpretation. According to the author, the learning process covers a broad spectrum of activities which can motivate both lookup and exploratory search. However, the investigative search activity is primarily an exploratory search, which “involve[s] multiple iterations that can take place over perhaps very long periods of time and that may return results that are critically assessed before being integrated into personal and professional knowledge bases“ (Gary Marchionini, 2006, p. 43).

Figure 2.2 Illustration that differentiates the iterative search from the exploratory search (Ryen W. White & Resa A. Roth, 2009, p. 21)

Ryen W. White & Resa A. Roth (2009) describe exploratory search as “[…] learning more about the topic of the search, understanding the nature of the document collection, and investigating browsing opportunities in real time as they occur during result examination” (p. 20). The authors specifically distinguish exploratory search from iterative search, deriving from the lookup search. The iterative search has a specific search target as can be seen on the left side of Figure 2.2. On the contrary, the lack of search target transforms the exploratory search strategy, compelling the information seeker to explore a bigger area, as illustrated on the right side of Figure 2.2. Each begins with a large information space that, following initial searches, are reduced to a smaller, more limited space. However, with exploratory search, the individual traverses many
different and interrelated information spaces while possibly searching for additional information or serendipitously discovering new spaces. This is not the case within a lookup iterative search because the searched information space is roughly defined by the initial search objective. The lack of a specific objective can also be defined as a “negative search” (Garfield, 1970) as suggested in Ryen W. White & Resa A. Roth (2009). For example, finding a research gap could be defined as a “negative search” and could trigger an exploratory search.

Several independently-studied states can be associated to exploratory search. For example, curiosity is defined as a state or an emotion which can be a source of motivation for exploration. Curiosity has been studied for a long time in psychology and several trends have existed along the years. It can be defined as a “driving force” and as an “impetus behind scientific discovery”. Curiosity is linked to exploratory behaviour since the early century. It is first experimented with animals where exploratory behaviour is described as the expression of the animals’ curiosity. Later, curiosity was related to human and more specifically to their intelligence as a factor and a motive for knowledge exploration (Loewenstein, 1994, p. 75). Then, the Pace’s model (2004, p. 343), distinguishes the specific and divergent curiosity. The first triggers well-defined goal-orientated behaviour while the second triggers ill-defined explorative behaviour. Curiosity can be seen as one factor motivating the information-seeker to enter whether a lookup search or an exploratory search. But, as noticed by Pace, curiosity is difficult to measure and do to provide satisfactory results.

Bates’ (1989) berrypicking model can also relate to exploratory search. Indeed, it is an information-seeking model, specifically developed for online systems, which followed the traditional information-retrieval models. It describes the search process as a succession of information-seeking steps where the information-need and the query evolve along the search process. The berrypicking model breaks away from the idea of a static retrieval from one query followed by its related set of results and introduces the idea of a step-by-step process where information is gathered bit-by-bit and where travelling toward the search goal is not linear. Bates also introduces the browsing behaviour as a succession of four steps. The information-seeker starts “glimpsing a field of vision”, followed by “selecting or sampling a physical or representational object from the field, examining the object and physically or conceptually acquiring the examined object, or abandoning it” (Bates, 2007). Bates’ first step of browsing reminds what can
be achieved by visualisation interfaces when providing information spaces’ overviews. Indeed, the visual overview naturally allows “glimpsing a field of vision”.

**Sensemaking** is defined as an activity or a behaviour which often involves a step where the information-seeker creates a representation of the information space. *Sensemaking*, like *curiosity* can be seen as a redundant activity being part of the exploratory search process. Indeed, the information-seeker moves into the information space and needs to make sense of unknown spaces (Ryen W. White & Resa A. Roth, 2009, p. 32).

One possible consequence of exploratory search is *serendipity*. It can be defined as a positive discovery resulting from a research which goal is not related with the discovery. In other words, a serendipitous discovery is new knowledge acquired while searching for unrelated information. The *precipitating conditions* of the serendipity process are the conditions which support and increase the chances of making a serendipitous discovery. Sensemaking, curiosity and exploratory search are part of the conditions that support serendipity (McCay-Peet & Toms, 2010).

Interactive information retrieval (IIR) is a research field which stems from information retrieval and information seeking. It pertains to possible means for retrieving information from the information-seekers’ perspective. There is a long continuum of research studies from the system-centred studies to the user-centred studies. IIR stands in the middle of the continuum with studies that evaluate the systems as well as the user’s interaction. However, IIR studies have focused slightly more on the systems and slightly less on the users’ behaviours (Diane Kelly, 2007, p. 10). The evaluation of information-seeking interfaces within an IIR context is challenging because of additional human factors, such as uncertainty, which makes the analysis more complex. Indeed, the *precision* and the *recall* indicators cannot be computed since the results cannot be labelled as right or false when it comes to sensemaking, learning or discovery (Wilson, Kules, Schraefel, & Shneiderman, 2010). Therefore, the traditional information-retrieval metrics are not measured in this project, in favour of other metrics built around the information-seeker’s behaviour and cognitive state.

The Figure 2.3 shows different research domains such as information retrieval, sensemaking and information visualisation with their respective focus such as interactive and cognitive information retrieval, berrypicking and exploratory search. The diagram clearly supports the idea of a multidisciplinary study, which might be defined
as an information visualisation study, with a special interest about the user’s exploratory search behaviour which can be situated on the behaviourally-side of an IIR study.

![Diagram showing exploratory search and other domains](image)

*Figure 2.3 The diagram shows where exploratory search is situated around the different research domains, models and theories (Ryen W. White & Resa A. Roth, 2009, p. 39).*

2.1.4 Information visualisation

2.1.4.1 The definition

Information visualisation is an interdisciplinary research domain that explores how to represent and interact with abstract data. It is a growing multi-domain field that includes computer science and aims to support human-cognition or behaviours. Chen (2005) defines information visualisation “as visual representations of the semantics, or meaning, of information. In contrast to scientific visualization, information visualization typically deals with nonnumeric, nonspatial, and high-dimensional data” (p. 12). This definition fits well this project’s use of information visualisation as it relates to representation of semantic information extracted from a collection of documents. It also relates to non-spatial and nonnumeric metadata (title, authors, description) that are transformed into a high dimensional document-term matrix. The challenges discussed by Chen at that time are still relevant. Finding visual interfaces’ measures of quality,
transforming complex analytical processes into simplified colourful images, designing one interface for several users’ prior knowledges, educations and trainings remain challenges that need to be considered during the design phase of every visualisation project.

2.1.4.2 The design studies

The design study is one type of information visualisation paper focused on the design of a visual representation meeting the needs of a specific domain problem. Most of the visual references cited in the section 2.2 - Design considerations can be considered as design studies such as the starSPIRE (Bradel, Wycoff, House, & North, 2015), the PATH (Goodale et al., 2013), the Refinery (Kairam et al., 2015), the Serendip (Alexander et al., 2014) which suggest visual search interfaces for improving and enhancing exploratory search. Munzner (2009) provides a framework guiding the visualisation designers writing up design studies, namely the four-level nested model whose schematic representation can be found in Figure 2.4. Several researchers, have measured the impact of data visualisation on information-seeking, however, the experiments do not distinguish the lookup tasks from the exploratory tasks, and the experimental user-tasks and metrics are mainly evaluated using the traditional information retrieval accuracy and efficiency measures (Hoeber & Khazaei, 2015; S. Liu et al., 2012; Wu & Vakkari, 2014).

![Figure 2.4 The four-level nested model provides a framework guiding the visualisation designer through four steps for developing a visual interface (Tamara Munzner, 2009, p. 922)](image)

2.1.5 Data analytics and exploration

Data visualisation and exploration are close concepts which share common history. For example, the literature shows some kind of general implicit acceptation toward visualisation enhancing exploration and discovery. This common acceptation is particularly strong in the data mining research field. For instance, Fayyad, Wierse, & Grinstein's (2002) influential book about visualisation and knowledge discovery
establishes a strong relationship between visualisation, exploration and knowledge discovery. Additionally, Foong (2002) concludes that visualisation is a tool improving the users’ exploration: “Visualization can be used […] as a stand-alone data-mining technique […] resulting in a more powerful and synergistic approach to data exploration and discovery” (p. 186). Another influential data mining paper assuming that visualisation is the right tool for data exploration is Keim (2002), concluding: “The ultimate goal is to bring the power of visualization technology to every desktop to allow a better, faster and more intuitive exploration of very large data resource” (p. 105). The same paper also defines Shneiderman’s (1996) well-known information-seeking mantra as the usual process for “visual data exploration”. As presented in those citations, data exploration is clearly stated as a result from the use of visualisation within the data mining analytics’ process. However, the term exploration is not defined and the exploratory behaviour is not evaluated.

Those papers are a good illustration of the data analytics context in which the terms exploration and discovery were used and associated to data visualisation. This trend fits the rise of the knowledge discovery in databases (KDD) and the visual data mining (VDM) research domains. However, it is interesting to notice that it seems to have a different meaning whether it is used in a data visualisation, data analytics or information-seeking context.

2.1.6 Data visualisation and exploration

The term exploration is also used within a controlled data visualisation typology. Indeed, Munzner & Maguire (2014) define exploration as one of the search action (explore) that takes place when performing a visualisation task, along with lookup, locate and browse (p. 54). The “explore” cell of the tab in Figure 2.5 is a visualisation mid-level user goal which is part of the visualisation user-task typology. The exploration action is part of the task abstraction which is one step of the design study as defined by Munzner (2008, p. 138). The task abstraction step is made of actions and targets. The three levels of actions added to one target define a user-task. The search action is part of the mid-level action of a user-task among with analyse and query. This typology therefore fits the information-seeking exploratory search definition where the search does not have a precise goal and does not know where to search nor which keyword to use.
Data visualisation has therefore widely been associated to a catalytic role for data exploration and discovery, however the question of whether visualisation supports or not exploratory search has not been clearly addressed in the literature.

2.1.7 The role of visualisation in digital repositories and libraries

Currently, the biggest digital libraries, like the Universal Digital Library illustrated in Figure 2.6, exclusively provide keyword-based search tools functioning as traditional web search engines such as Google. In both cases the user is confronted to a search bar which only accepts keywords. Therefore, the user’s search goal needs to be formulated with keywords, which generally requires both to have a precise goal and sufficient knowledge about the search context. A large majority of digital libraries have complied with the same system, even the smaller ones. Sometimes, supportive visualisations are
provided beside the lookup search interface, as a complementary tool for refining the search through filtering. The Springer’s AuthorMapper tool is one example of implemented hybrid search interface, as illustrated in Figure 2.7. Several visualisations such as a word-cloud, horizontal bar charts showing the countries, the institutions, the authors, the publication type and the authors’ map are used as supportive tools impacting dynamically the list-view. The core system remains the keyword-based search interface which can be used as a standalone search tool, whereas the small-multiple visualisations are dependent from the list-view.

![Figure 2.7 The Springer’s AuthorMapper: http://authormapper.com/search.aspx?q=visualisation (accessed the 16/08/2016)](http://authormapper.com/search.aspx?q=visualisation)

Fast & Sedig (2006) designed the Interactive Visual Environments (INVENT) framework which defends the use of information visualisation as a catalytic tool for reconceptualising the digital libraries’ environments. In the authors’ point of view, information visualisation supports the transformation of information into knowledge and

1 The Springer’s AuthorMapper: [http://authormapper.com](http://authormapper.com)
support the information-seeker into the digital information space. The shift, from physical to digital libraries mainly focused on the technical aspects of digitising and indexing documents, which facilitated access to information. However, the lookup interfaces are not useful yet when it comes to exploring and creating knowledge. Information visualisation is the right tool for developing visual interfaces and helping digital libraries in completing their transition (Fast & Sedig, 2006).

Several papers specialised in the digital libraries research field, strongly support the use of visualisation in digital libraries. Merčun & Žumer (2010) for example raise the information discovery, exploration and serendipity issues in today’s digital libraries. The paper’s statement however remains high-level and does not detail whether the visualisation should rather apply on the library as a whole or solely on the results’ space. Moreover, the authors make suggestions about specific visualisation design and encoding illustrated with subjective comments.

Data visualisation and exploration have already been associated in many different ways. In particular, researchers often try to understand the role of visualisation in information discovery and serendipity. Hinrichs et al. (2015) conducted a case study about the role of text mining and information visualisation in the exploration and discovery processes when searching a large scale historic document repository. The feedbacks gathered along the study are promising and show that the combination between text mining and interlinked information visualisations (tag cloud, location cloud, histograms, maps) enhances potential discoveries. However, the feedbacks are subjective and only highlight a potential enhancement. Moreover, the suggested visualisations are very analytical and intended to experts.

On the one hand, the literature tends to model the information-seekers’ behaviour, understanding better and better their search strategies and cognitive processes. On the other hand, there has been a long-lasting and global intuition towards visualisation’s promises for improving exploratory search. Numerous independent visual tools and interfaces were developed and loosely validated. This study therefore aims at better understanding the role and the impact of visualisation within the context of exploratory search and digital libraries.
2.2 Design considerations

This section is the continuation of the literature review, with a special emphasis on similar design studies from which design considerations are derived. Visualisation has been used in many different ways as a tool to support information-seekers’ search process within digital repositories and digital libraries. Several aspects of visualisation, as well as several aspects of the search task have been studied. The chapter is organised as a sequence of six high-level design considerations each of which is derived from several visual search interfaces’ studies and low-level considerations. The targeted papers usually apply on textual digital documents such as books, journals, articles, news or emails. The scope therefore excludes the web and the media such as websites, music, pictures and visual art in general. When possible, experimental designs are also reviewed from the literature in order to derive some experimental considerations.

2.2.1 The information space

Exploring a limited space of items resulting from a self-formulated query is fundamentally different than exploring an unknown space of items. The literature specialised in digital libraries often tends to focus on the space of results produced by the information-seeker’s query instead of focusing on the entire collection of documents. The user is therefore often limited to a sub-space of results, of which the scope is not fully controlled. Indeed, the workflow from the query to the results can be complex and opaque (Ruotsalo, Peltonen, et al., 2015; Tablan, Bontcheva, Roberts, & Cunningham, 2015). Whether visual or not, the process is derived from the traditional keyword-based search-query systems which only display the results from the users’ queries. It has proved its efficiency in regards to the precision and the recall, in situations where the information-seekers’ goals are clearly defined and formulated into a keyword-based query. This could be defined as a result-based visualisation, contrary to the library-based visualisation which would display the entire information space. For instance, the SciNet and the Refinery search interfaces ask users to enter at least one keywords in order to start the search process, even though the following steps are mainly visual (Kairam et al., 2015, p. 305; Ruotsalo, Peltonen, et al., 2015). They therefore include visualisations which might support exploration of the limited space of results. However, the visualisations are limited to users in possession of a goal which can be formulated

---

2 The precision focuses on the relevant selected items among all selected items resulting from the user’s query, whereas the recall focuses on the relevant selected items among all relevant items.
into keywords. In this context, the main shortcoming is twofold. First, the users without goal or way to formulate it cannot enter the search and cannot profit from the potential visualisations provided by the search interfaces. Secondly, the users entering the search can only visualise the limited space of results defined by their own search-query. Those constraints are avoided by providing the information space’s overview as suggested by Shneiderman (1996).

Similarly, the *adaptive VIBE* visual interface shows documents related to the users’ present and past queries at the same time. Every query generated from the user are indeed stored inside a user model. The visualisation is therefore a result from both a query and a user model. However, it does not display the entire collection of documents nor allow for navigation within an undefined information space. Indeed, the space has to be scoped with a query input (Ahn & Brusilovsky, 2013).

Therefore, the ExploViz search interface should provide an initial overview of the entire collection of documents, in order to support the users in building a competent mental model of the underlying information space and to facilitate navigation across the space’s items (Shneiderman, 1996, p. 339).

---

DC 1: Provide a visual overview of the a full-scope collection of underlying documents.

DC 2: Provide an initial visualisation without requesting users’ input.

2.2.2 Spatialisation of metadata

The spatial metaphor of text documents was introduced in the nineties by Skupin, A. & Buttenfield, B. P. (1996) in a paper presenting the use of a document-term matrix followed by a dissimilarity matrix and a multidimensional scaling for dimensionality reduction and documents’ spatialisation. Today, the basic concept of the spatialisation technique remains the same and can be used in combination with other natural language processing algorithms and similarity metrics. In addition to the spatial metaphor, Tobler’s first law of geography introduces the relation between similarity and spatial closeness: “everything is related to everything else, but near things are more related than distant things” (Tobler, 1970, p. 3).
Figure 2.8 The 3D Vase Museum is a 3 dimensional scatterplot encoding the years and wares. The third dimension is not informative. It is rather used for displaying the room and motivate the user navigating in a 3D space (Shiaw, Jacob, & Crane, 2004)

Figure 2.9 The ActiveGraph is a scatterplot that provides the user with 6 possible encodings. This illustration shows the publication years in abscissa and the authors both in ordinate and in colour (Marks, Hussell, McMahon, & Luce, 2005)
Following those spatial concepts and techniques, several spatial visualisations were developed for displaying large collections of digital documents on a two dimensional plane. It is used as a way to order and position books virtually which can enhance digital libraries’ sensemaking. Early visualisations of digital libraries used scatterplots for encoding the information space. For example, the 3D vase museum and the ActiveGraph use the documents’ metadata as direct information for the scatterplots’ axes. Either two or three dimensional, with linear or logarithmic scales, the scatterplots display the documents in a rather analytical way. As can be seen in Figure 2.8 and Figure 2.9, the axes are made visible on purpose so that the users can either search books based on the metadata or navigate inside the scatterplot. Even though the collection can be displayed as a whole, the spatialisation is limited to a few numerical or textual metadata. Two spatially close documents are therefore similar to the extent of their limited encoded metadata, resulting in a restricted spatialisation. When only two metadata are encoded, as the 3D Vase Museum in Figure 2.8, the spatialisation information is limited and the third dimension is not useful. On the contrary, when six metadata are encoded, as the ActiveGraph in Figure 2.9 (x, y, z, size, colour, shape), the visualisation can become difficult to approach. It therefore increases the learning curve and decreases intuition and motivation to use the visualisation. This is the main limitation of those early visualisations (Marks et al., 2005; Shiaw et al., 2004).

![Diagram](image)

*Figure 2.10 The spatialisation step from a pre-processed tree constructed from a thesaurus and the final map where the tree’s hierarchy is respected. This step is part of the Hierarchical Spatialisation Algorithm (HSA) presented in (Hall & Clough, 2013)*

The PATHS project aims at developing a search interface improving exploration and navigation. It started with a motivation to use the spatial metaphor for displaying digital documents on a 2D plane, as a map. The authors first introduced the Hierarchical Spatialisation Algorithm (HSA) which is a novel algorithm providing for hierarchical spatialisation. The workflow consists in pruning a tree made of a thesaurus and assigning
each document, from the digital collection, to one leaf. The spatialisation is then performed using a multidimensional scaling (MDS) based on the traditional TF-IDF document-term matrix and the bag-of-words vectorisation technique. The MDS is adapted for preventing any spatial outliers and any overlapping of topics at both the parent and child levels, as can be seen in Figure 2.10. This spatialisation technique has the advantage of displaying the whole collection of documents on a 2D plane, allowing for novice users to overview the entire digital space (Hall & Clough, 2013). However, it asks for a hierarchical thesaurus as input, which prevents from spatializing unlabelled collections of documents. Also, the HSA asks for a perfect separation between documents. Each document must therefore relate to a unique topic, which seems unrealistic. This constraint echoes on the design because the MDS is specifically adapted for preventing topics from any spatial overlapping. Those two strong constrains are avoided when using the well-known topic modelling Latent Dirichlet Allocation (LDA) algorithm.

![PATHS exploration tab which displays a high-level map of the themes contained in the entry thesaurus. The parent nodes are not overlapping among each other and the child nodes are well contained within its corresponding parent (Goodale et al., 2013).](image)

The spatialisation algorithm is used in the PATHS project for developing the “map” tab of a multi-view search engine as illustrated in Figure 2.11. The tab is specifically developed for exploration purposes and it is specifically evaluated in comparison to a
previous non-visual version of the same search interface. However the experiment is entirely based on qualitative feedbacks and do not provide conclusive remarks except that novices are more likely than experts to perceive the “map” tab as a useful exploration tool (Goodale et al., 2013). The paper would have profited from a quantitative experiment. Moreover, preventing the topics and the documents to overlap and showing topics without representing the documents are the two main shortcomings of the PATH’s map tool.

Later on, the reference map is used for organising the search results. It can be defined as a spatial map where each region corresponds to a topic. The Figure 2.12 outlines the difference between the reference map visualisation (on the left) and the results encoded in an orange network and displayed on top of the reference map (on the right). The map supports the users’ mental representation of the space and enhances sensemaking. It remains on the background and it adapts dynamically to the queries by changing the areas’ size depending on the distribution of the resulting documents’ location. However, the search interface is designed toward the results’ space because the documents are only accessible within the network after entering a search query. In other words, the user cannot start exploring the information space from the global overview provided by the reference map. Another limitation comes from the use of a document graph to display the search results which asks for important resources and calculation time. Finally, as a substitute for evaluation, the authors provide a step-by-step tutorial explaining how to use the application.
Another form of spatialisation can be seen in Figure 2.13. The starSPIRE spatialized search interface is an example of visualisation specifically designed for sensemaking, semantic interaction and information retrieval. Indeed, the spatialisation is used for displaying the results originating from the user’s keyword-based query-search. Local sensemaking is therefore potentially enhanced around the results’ space. However, global sensemaking about the entire collection of underlying documents, is made difficult. A semantic interaction is provided by interpretation of the users’ interaction with the system. For instance, when users move nodes and reshape the space, the underlying weighting scheme changes accordingly. This feature is similar to Ruotsalo et al.’s (2015) SciNet where the users’ feedbacks are used back into the model for future queries.

The spatialisation technique is of special interest because it allows to provide for the entire information space’s visualisation as an entry point for exploration and sensemaking. The latter is expected to be static in order to avoid incorporating noise and unnecessary complexity within the evaluation process. Moreover, it should be possible to use a model which learns a hierarchy from the collection of documents and it should be possible to derive more than one topic from each document. Finally, the reference map concept is kept in mind and shifted toward a simpler static background concept, as it will be further explained in chapter 3 - Design and implementation.

DC 3: Spatialize the collection of documents on a 2D plane.
DC 4: Suggest a hierarchy from the collection of documents.

DC 5: Derive several topics from each document.

DC 6: Add reference map to support the user’s mental representation.

2.2.3 Topic modelling

The Refinery search engine, as illustrated in Figure 2.14, is designed for supporting the associative browsing strategy. It can be defined as one exploratory search type specifically used by information-seekers who browse information from one item to the next similar items. The visualisation is based on a network dataset. It therefore contains information of relatedness between each node. The search entry point is a keyword query provided by the information-seeker. This visualisation is a good example of search interfaces that support other exploration strategies based on other type of entry dataset. The authors define this strategy as a bottom-up exploratory search rather than the more traditional Shneiderman’s (1996) top-down overview first mantra. The evaluation is based on a subjective questionnaire preventing from measuring the respondents’ behaviours in favour of their perceived experience about their interaction with the tool (Kairam et al., 2015).
The Grisham word-cloud tab allows the user to explore among the keywords forming a topic. The bigger the word the more significant it is inside the topic (Grant et al., 2015).

The Grisham visual search interface leverages the Latent Dirichlet Allocation (LDA) topic modelling. Each document from the collection is pre-processed and transformed into a combination of topics and its corresponding probability. The search engine is also based on a user model which consists in a vector of weights where each weight is associated to one topic found in the corpus. The latter can be can be manually personalised by the information-seeker. Additionally, the search interface provides for a word cloud visualisation (Figure 2.15) allowing for each topic’s keywords exploration and a doughnut visualisation showing the proportion of all topics composing a document. As highlighted by the authors, the topic-based search interface asks for topic labelling, which is an additional cognitive load for the user and a possible drop-down reason for exploration. Even though the word-cloud visualisation probably facilitates the topic labelling, it is believed that it would not be an efficient visualisation for presenting one nor several topics of a digital library. Finally, the paper does not provide an evaluation nor a comparison of its suggested search interface. It is therefore difficult to evaluate the efficiency of such visualisation for searching a collection of documents using a topic-based search approach (Grant et al., 2015).
The Serendip search engine is also based on the probabilistic LDA topic model. It aims at enhancing serendipitous discovery through multiple and interconnected views about the topics and the text corpus. The CorpusViewer shows the main matrix of probabilities resulting from the LDA algorithm. The probabilities of the document-topic matrix are encoded as coloured glyphs. The matrix can be reordered, which fosters serendipitous discovery. Additionally, the TextViewer displays the content of a selected text with a line graph on the side showing the strength of the topics inside the document. Finally, the RankViewer displays the topics’ terms as an inverted bar chart where each term is represented as a grey slice. The three views can be seen in the Figure 2.16. Terms can be highlighted with a colour encoding. The authors mainly focused on visualisations that express the direct results from the LDA, namely the document-topic matrix, the topics’ terms and their proportions inside each topic. Even though the search does not require an entry keyword nor a query, the visualisations are only centred around keywords and topics. A document is showed as a collection of terms and its resulting probabilities. This can be seen as a limitation for users who prefer navigating and exploring books instead of words (Alexander et al., 2014).
Figure 2.17 The LDAvis (Sievert & Shirley, 2014) is a visual ad topical search engine. It provides for multiple visualisations mainly representing the topics and the terms.

Similarly, the LDAvis topical search interface, is designed for exploring topics underlying large collections of textual documents. Once again, the topic is the central object of the search instead of the book. This makes the LDAvis more analytical and syntactical than traditional search engines. Topical search tends to be designed for experts or advanced users rather than novices. Indeed, the user-tasks are more specialised and focused on keywords and topics instead of covers and positioning, preventing from casual browsing or open-ended exploration. However, contrary to the Serendip interface, spatialisation is used for displaying the topics on a plane, as illustrated in Figure 2.17. The user can also interact with the main underlying spatialisation parameters, which fosters serendipitous discovery. The main shortcoming of LDAvis therefore lies in its limited topical representation with makes it a tool mainly targeted to an advanced audience. The user-friendly input dropdown list and sliders make it easier to interact with but the users cannot not access the documents directly.
The Mimir semantic search interface is specifically designed for performing precise search. Besides, it also aims at improving discovery and exploration by the use of visualisation. The tool therefore offers a combination of lookup search and exploration of the resulting space from the query. As outlined in the paper, exploration is supported by a word-cloud, similar to Grisham, and a co-occurrence terms matrix, displayed in Figure 2.18 (Tablan et al., 2015). The matrix is supposed to facilitate terms’ associations and discovery. The visual and exploratory aspect of the tool are not evaluated in favour of the indexing and search efficiency.

To sum up, several visual search interfaces leverage topic modelling as a tool for exploring the documents’ topics and enhancing serendipitous discovery. The visualisations are generally incorporating the topics’ keywords inside bar charts (LDAvis, Serendip), probability matrices (Serendip), co-occurrence matrices (Mimir), word-clouds (Grisham, Mimir) or highlighting the keywords into the text directly (Serendip). Moreover, spatialisation is also used for displaying the topics on a plane (LDAvis).

DC 7: Support hierarchical analysis and provide a starting point for deeper analysis with topic modelling.
Display the collection of documents in combination with the topics resulting from the topic modelling.

2.2.4 Nice, easy and intuitive to use

The Bookfish digital library targeted for children, as illustrated in Figure 2.19, is a good example of a visual search interface that does not incorporate a keyword search tool but rather some sliders that can be positioned by the users in order to refine the search. The sliders allow the user to provide a structured query to the system based on six book and reading preferences such as book’s difficulty, fantasy level and seriousness level. It therefore is a good mix between a search tool and a recommender system. Even though no keyword is needed for the entering the search, it does not provide an overview of the entire space and the navigation is only permitted through the six sliders. Therefore, the query system is strongly limited, as well as the library’s sensemaking opportunities. Indeed, exploration is only possible around the suggested axes but not around the books’ space (Pearce & Chang, 2014).
The SIZL (“Searching for Information in a Zoom Landscape”) is a recent interface that has the particularity of offering a 2.5D visual environment. The tool shows both a 3D global view of the underlying digitalized items and a 2D traditional view of the documents if the user wishes to zoom into the selected document. The tool is an interesting hybrid as it allows for global exploration into a 3D space of documents as well as entering keywords for the traditional search tasks. A comparison between SIZL and a traditional file explorer system is performed. The user-tasks are three lookup search questions with varying difficulties. The lookup search tasks performed in SIZL show similar levels of accuracy and significantly lower amounts of time to perform the tasks, in comparison to the file explorer baseline (Grierson et al., 2015). The evaluation, although quantitative, is limited to a small scope of documents and lookup search tasks only, instead of exploratory search tasks. Moreover, the SIZL, similarly to the Bookfish library, displays thumbnails which prevents from displaying too large volumes of documents.
Figure 2.21 The SciNet (Ruotsalo, Jacucci, et al., 2015) is a spatialisation shaped as a radar where the user is located in the centre of the visualisation. The user can interact with the system by selecting keywords and dragging them toward the centre of the circle.

The literature tends to show that some complex modelling-based researches also try to tackle the lack of exploratory search support in web and digital libraries search engines. The user-centred “interactive intent modelling” focuses on the fact that users regularly switch from lookup to exploratory behaviour while performing a complex and iterative search task. It therefore tries to anticipate the user’s state based on his feedbacks and interaction with the system. The SciNet visualisation illustrated in Figure 2.21 was created accordingly. The authors performed a rigorous evaluation based on a comparison against a traditional list-view baseline. On the one hand, two novel metrics try to capture the levels of exploration and discovery by measuring the type of interactions and type of information retrieved (novel versus obvious). On the other hand, precision, recall, F-measure and answers’ scores are provided by two post-doctoral experts who also designed the search tasks. It is believed that avoiding subjective and manual evaluation
of the users’ answers would be of benefit to the evaluation in order to reduce the risk of potential bias and to reduce the results’ complexity of interpretation (Ruotsalo et al., 2013, p. 1762,1764; Ruotsalo, Jacucci, et al., 2015, p. 88).

![Diagram of Adaptive VIBE](image)

Figure 2.22 The Adaptive VIBE (Ahn & Brusilovsky, 2013, p. 1144). This shows the resulting screen of the user’s “Nuclear

Similarly, the *adaptive VIBE* visual search engine is a spatial visualisation where the users’ query is presented as a point of interest on a plane, surrounded by similar documents. Additionally, the users’ relevant past queries are added within the visualisation and reshape it. As can be seen in the Figure 2.22, the user entered the query “Nuclear Weapon”. The two words form two distinct points of interest. The purple points represent the related past queries. The documents are then positioned based on their respective similarity to the other documents and points of interest. The authors specifically outline that the *adaptive VIBE* search tool is designed for experienced users. The visualisation is therefore based on both a query (issued by the information-seeker) and an underlying user model which is stored and constantly evolving. The interface also provides for a selection tool which is a separate visualisation using different encoding. The evaluation is based on a comparison between the authors’ *adaptive VIBE* and a list view baseline called *TaskSieve*. Because the designed tool is rather new and
complex, the experiment starts with an introduction and a 50 minutes training. This is a constraint limiting the experiment to a small group of advanced users set up in a laboratory configuration. Indeed, the learning requirements are too complex to be used by a casual and untrained audience. The two following search activities combine exploratory search and fact findings. Finally, the comparison is mainly based on several versions of the precision and recall metric. Each document was therefore previously tagged as relevant or not for each corresponding search task (Ahn & Brusilovsky, 2013, p. 1144). The adaptive VIBE is believed to target experienced information-seekers with a high motivation to learn the system. Moreover, the interface’s adaptiveness to past queries can be seen as a lack of transparency which can potentially cause uncertainty and impair the exploratory search. The within-group evaluation is rigorous and quantitative but it required carefully selected subjects and a complete laboratory configuration.

Figure 2.23 The Bohemian Bookshelf (Thudt, Hinrichs, & Carpendale, 2012) is a digital library focusing mainly on curiosity, exploration, multiple interconnected views and a playful interface. The picture shows the five views available to the user when searching a book. Each view is the expression of a specific metadata (cover, number of pages, authors, keywords or year of publication).

The Bohemian bookshelf is a multi-view visualisation that aims at enhancing exploration, discovery and serendipity through a search interface of five interlinked visualisations. Each visualisation shows a book’s specific metadata such as the author, the cover, the number of pages, the tags and the year of publication. The authors also
provide some design considerations for increasing serendipity such as “*multiple visual access points*”, “*highlighting adjacencies*”, “*flexible visual pathways*”, “*enticing curiosity*” and “*playful exploration*”. The multiple visual access points can be seen as a double-edged sword because multiple coordinated views can potentially require additional cognitive attention (Thudt et al., 2012).

**DC 9:** Avoid coordinated visual displays.

**DC 10:** Do not integrate user intent into the spatialisation.

**DC 11:** Maximise transparency to minimise uncertainty; user is in complete control.

**DC 12:** Motivate curiosity with a playful interface.

**EC 1:** Focus on objective and quantitative metrics to support comparison and further experimentation.

**EC 2:** Perform the evaluation against state-of-the-art – traditional search interface.

**EC 3:** Foster an online experimental design targeted toward a casual audience.

2.2.5 “Overview first, zoom and filter, then details-on-demand”

*Figure 2.24 The WikiGalaxy is a 3D visual interface displaying around 100 000 Wikipedia articles as a galaxy. The Home tab shows an overview of the galaxy. Each colour corresponds to one theme. Retrieved from http://wiki.polyfra.me/ (accessed the 12/07/2016)*
Shneiderman's mantra can be considered as one early and influential design consideration (1996). The mantra firstly stipulates that an overview of the collection should be available in visualisations. Among the previously cited visual search interfaces using information visualisation as a tool to enhance exploratory search, only a few propose an overview of the collection. The starSPIRE, Grisham, Bookfish, SciNet, adaptive VIBE only display the space of results following the users’ queries. On the contrary, the Serendip, PATH, LDAvis and Bohemian bookshelf search interfaces provide an overview of the underlying information space. This is made possible because the data is aggregated into topics or because the underlying information space is small.

A good example of visual information space overview is the web-based Wikigalaxy interface\(^3\) which displays around 100 000 Wikipedia articles as a 3D galaxy containing nebulae. As can be seen in the Figure 2.24, the entire collection of articles can be visualised in the Home mode. The colour encoding relates to high-level themes, even though it is not clearly defined. However, the Home mode is fixed and the users cannot zoom inside the galaxy. On the contrary, the Map and Fly modes allow for navigation into the Wikipedia’s galaxy. However, moving around the articles is not easy because of the size of the space and also because of the movements’ directions limitations. The Map and Fly modes can be used for finding details about relations between articles as well as for displaying a specific article’s content. A search box can also be used for tracking articles. The Figure 2.25 shows the interface when the Belfast article is selected. The links can be clicked in order to move to the other related articles. Although this is a beautiful piece of design, in term of practical usability, it is very difficult to orientate in a 3D space and the users can easily get lost.

\(^{3}\) Wikigalaxy: [http://wiki.polyfrea.me/](http://wiki.polyfrea.me/) (accessed the 12/07/2016)
Figure 2.25 The WikiGalaxy is a 3D visual interface displaying around 10000 Wikipedia articles as a galaxy. The fly mode orients the galaxy for improving navigation inside it. When an item is clicked, the related articles are highlighted with white links. Retrieved from http://wiki.polyfra.me/ (accessed the 12/07/2016)

Figure 2.26 The Overview second version as illustrated in (Brehmer, Ingram, Stray, & Munzner, 2014). The visualisation of interest is the spatialisation scatterplot on the upper right part of the picture. The scatterplot was finally remove from the final Overview tool because of the inability to access the documents’ contents.

Brehmer et al.’s (2014) design study introduces the Overview interface, which allows users to explore large collections of untrusted and unclassified documents. The Overview tool aims at analysing the underlying documents’ content. It also aims at generating and verifying hypotheses, exploring and summarizing documents. The Overview tool used a 2D scatterplot in its early versions (V2 as illustrated in Figure 2.26) but it was finally removed because the scatterplot was not useful for showing the documents’ content. Consequently, the scatterplot was not adopted nor evaluated in the Overview case study. One limitation of the Overview scatterplot therefore lies in its inability to show details about the documents it contains. Moreover, although this
specific visualisation is of special interest, it was not tested nor evaluated, which makes it difficult to support.

DC 13: Present an overview from which it is possible to zoom into the space.

DC 14: Contextualise the search within the underlying information space.

DC 15: Avoid 3D representations to prevent from getting lost into the information space.

DC 16: Use space to encode semantics and colour to encode hierarchy and topics.

DC 17: Add a card on hover in order to access the documents’ content.

2.2.6 Interactivity

As outlined in Pace’s model (2004, p. 343), curiosity is a key factor determining whether an information-seeker enters a lookup search or an exploratory search. Especially the dersive curiosity generates an exploratory search behaviour.

![Jigsaw document cluster feature](image)

_in order to go beyond the traditional lookup search provided by the keyword search interfaces, the Jigsaw interface makes use of automated text analysis and interactivity between a visual interface and the information-seeker. The user can interact with the slider which is connected to the visualisation, as illustrated in Figure 2.27. This triggers curiosity and serendipity by providing document similarity and document clustering_
adaptive visualisations. However, several limitations apply to Jigsaw. First, the
document similarity shows similar documents relative to one selected document instead
of relative similarity of documents among others (Figure 2.28). Second, the two features
are therefore separated and not interconnected which requires the user to switch tab
when using one or the other. Finally, the paper does not provide a formal evaluation but
instead it provides explanations about how to complete two investigative search
scenarios (Gorg et al., 2013).

![Jigsaw similarity feature where each rectangle is a document and the similarity to the chosen document is encoded into the rectangle’s opacity (Gorg et al., 2013).](image)

Figure 2.28 Jigsaw similarity feature where each rectangle is a document and the similarity to the chosen document is encoded into the rectangle’s opacity (Gorg et al., 2013).

DC 18: Add interaction to the search interface for increasing the information-seeker’s curiosity.

2.3 Summary

In summary, the reviewed design studies demonstrated visual search interfaces aiming
at supporting exploration, sensemaking, curiosity and playfulness. Topic modelling is
often used and displayed as such (Alexander et al., 2014; Grant et al., 2015; Hall &
Clough, 2013; Sievert & Shirley, 2014; Tablan et al., 2015), which makes the interfaces
difficult to approach and designed for experts and for analytical purposes. Suggested
visualisations are also regularly presented as spatial representations based on
spatialisation techniques (Bradel et al., 2015; Goodale et al., 2013; Grierson et al., 2015;
Nocaj & Brandes, 2012; Ruotsalo, Peltonen, et al., 2015; Sievert & Shirley, 2014), which
often do not present overviews of the entire underlying information spaces. Moreover,
the evaluations are often suggested for future works (Smith, Hawes, & Myers, 2014, p. 77). If not, evaluations are either qualitative (Kairam et al., 2015, p. 307) or quantitative
and relative to a baseline (Lin et al., 2015, p. 163; Ruotsalo et al., 2013, p. 1762).
Additionally, the respondents’ cognitive load might be computed while performing search tasks, as in (Y. Liu, Barlowe, Feng, Yang, & Jiang, 2013, p. 32). The resulting design considerations reflected from the reviewed design studies are illustrated in Table 2.1. They are specifically informing the ExploViz interface design discussed in the next chapter.

<table>
<thead>
<tr>
<th><strong>The information space</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DC 1: Provide a visual overview of the a full-scope collection of underlying documents.</td>
<td></td>
</tr>
<tr>
<td>DC 2: Provide an initial visualisation without requesting users’ input.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Spatialisation of metadata</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DC 3: Spatialize the collection of documents on a 2D plane.</td>
<td></td>
</tr>
<tr>
<td>DC 4: Suggest a hierarchy from the collection of documents.</td>
<td></td>
</tr>
<tr>
<td>DC 5: Derive several topics from each document.</td>
<td></td>
</tr>
<tr>
<td>DC 6: Add reference map to support the user’s mental representation.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Topic modelling</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DC 7: Support hierarchical analysis and provide a starting point for deeper analysis with topic modelling.</td>
<td></td>
</tr>
<tr>
<td>DC 8: Display the collection of documents in combination with the topics resulting from the topic modelling.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Nice, easy and intuitive to use</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DC 9: Avoid coordinated visual displays.</td>
<td></td>
</tr>
<tr>
<td>DC 10: Do not integrate user intent into the spatialisation.</td>
<td></td>
</tr>
<tr>
<td>DC 11: Maximise transparency to minimise uncertainty; user is in complete control.</td>
<td></td>
</tr>
<tr>
<td>DC 12: Motivate curiosity with a playful interface.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>“Overview first, zoom and filter, then details-on-demand”</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DC 13: Present an overview from which it is possible to zoom into the space.</td>
<td></td>
</tr>
<tr>
<td>DC 14: Contextualise the search within the underlying information space.</td>
<td></td>
</tr>
<tr>
<td>DC 15: Avoid 3D representations to prevent from getting lost into the information space.</td>
<td></td>
</tr>
<tr>
<td>DC 16: Use space to encode semantics and colour to encode hierarchy and topics.</td>
<td></td>
</tr>
<tr>
<td>DC 17: Add a card on hover in order to access the documents’ metadata.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Interactivity</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DC 18: Add interaction to the search interface for increasing the information-seeker’s curiosity.</td>
<td></td>
</tr>
</tbody>
</table>

*Table 2.1 Overview of the design considerations informing the visual ExploViz search interface*
3 DESIGN AND IMPLEMENTATION

This chapter discusses the design and implementation of the ExploViz interface, which is a visual spatialisation designed specifically for contrasting the non-visual state-of-the-art search interfaces in digital libraries. First the section 3.1 introduces and details the ExploViz interface design following a problem-driven design study framework as defined by Munzner (2008, p. 138). Inside this section, the nested-model’s four steps are presented in combination with the multi-level typology’s what, why and how. Then, the section 3.2 discusses the ExploViz interface implementation from the data collection, profiling and cleansing to the data transformation using topics’ modelling algorithm and dimensionality reduction for spatialisation. The section 3.3 concludes the chapter.

3.1 The ExploViz interface design: a visualisation design study

As previously presented, the Munzner's (2009) four-level nested model is used as a framework for developing the visual tool. In addition to the model, the following design study also complies with Brehmer & Munzner's (2013) multi-level typology, which is fully consistent with the four-level nested model. The Figure 3.1 shows an overview of the visualisation framework used for designing both the visual interface that is named ExploViz and the task-based experiment, highlighted in yellow. The design of the visual interface follows the information visualisation problem-driven design study. The four main steps of the nested model are detailed in two main papers (Meyer et al., 2012; Tamara Munzner, 2009). The following section details the four nested steps in combination with the multi-level typology and aims at justifying the ExploViz interface and the user-tasks used in the experiment.
3.1.1 Domain problem characterisation

The domain problem characterisation is mainly derived from the literature review. There is indeed a general acceptance about the lack of support when users enter an exploratory search in digital repositories (Aletras, Baldwin, Lau, & Stevenson, 2015; Brehmer et al., 2014; Gary Marchionini, 2006; Hoeber & Khazaei, 2015; Wildemuth & Freund, 2012; Wilson et al., 2010). The exploratory search is well-defined since it was the central subject of the ACM SIGIR workshop in 2006 (White, Muresan, & Marchionini, 2006). Years later, Wildemuth & Freund, (2012) summarized exploratory tasks as search tasks that “focus on learning and investigative search goals; they are general (rather than specific), open-ended, and often target multiple items/documents; they involve uncertainty and are motivated by ill-defined or ill-structured problems; they are dynamic and evolve over time; they are multi-faceted and may be procedurally complex; and they are often accompanied by other information or cognitive behaviours,
such as sensemaking” (p. 1). The definition outlines the main characteristics on which the present project focuses. It also informs the visual interface’s design.

In addition to the clearly defined gap in the literature, a case study was undertaken in order to inform and enrich some aspects of the domain problem characterisation. The case study pertains to the Digital Content Explorer (DICE) which is a visual search interface designed and implemented by Cyberlibris (Figure 3.2).


Cyberlibris is a medium-size digital library offering mainly academic contents. Understanding why the DICE was created at Cyberlibris and why it is used, is a good way to document and illustrate this project’s research question⁴. The two main contributors of the DICE’s creation were interviewed about the project’s origins and development. Globally, the conversations enlighten the author about the use of visualisation, and more particularly the spatialisation technique. It tells that the DICE is the intuitive result of a five years’ data analysis project originating from a frustration about the incapacity to present the readers with a clear view of the tremendous amount of metadata generated from the operational use of the digital library’s platform and

---

⁴ The case study, in line with this project’s research question, was initiated and written the same year by the same author as part of another academic assignment, namely ASTC2201: Case Studies in Computing, taught by Andrew Hines.
collected by Cyberlibris. The managing director has envisioned a visual space of the library’s information space in order to enhance navigation and collaboration among the books and the readers. Even though the DICE is not specifically designed for exploration, it is believed that it supports exploratory search. This is the reason why the suggested visual search interface resembles the DICE, in some respects. The case study also outlines the organisation of the five years DICE visualisation project, which first was a data analytics project. The search interface’s development process was mostly based on trials and errors as well as intuition, rather than task-based experiments and usability tests. The present visualisation study therefore also contributes to formally test empirically the impact of spatialisation on the readers’ exploratory search process.

3.1.2 Data and task abstraction design – What and Why?

The *data and task abstraction* is the second step of the four-level nested model (Meyer et al., 2012; Tamara Munzner, 2009). It is similar and often associated with the *What* and *Why* steps of the Brehmer & Munzner's (2013) typology, as outlined in Figure 3.1. The data and the task abstractions are crucial in a problem-driven project because they shape both the search interface’s visual encoding and the task-based experimental design. Precisely, the design study task abstraction is used to explain the data transformation resulting to the visual interface and to explain the tasks the users can perform with the visual interface. The same typology will be used later for justifying the user-tasks that are designed for the online experiment. The present section therefore highlights how the domain problem is abstracted into tasks that both transform the initial data into derived visual data and inform about the possible tasks that the interface offers to the users.

3.1.2.1 The tasks – Why?

“The why part of our typology, […] allows us to describe why a task is performed, and includes multiple levels of specificity, a narrowing of scope from high-level (consume vs. produce) and mid-level (search) to low-level (query).” (Brehmer & Munzner, 2013, p. 2378). Following this definition, the three levels can be adapted into the ExploViz’ three level tasks.

High-level – *consumption*: The ExploViz aims at providing the users with means of information consumption. This is believed to be motivated by either *discovery* or *enjoyment*. Indeed, searching a book can be the result of any personal interest in reading
a book. The users may also want to infer general information about the digital library, such as the distribution of books or the clusters of similar books.

Mid-level – search: The search actions are defined depending on whether the target and the location are known or not. The ExploViz focuses on the “explore” action which fits well the exploratory search as defined in information-seeking. Indeed, either the location nor the target are none when entering an exploratory search. This is central in the design because it fosters the use of visual tools that are believed to increase the users’ motivation for exploration, such as spatialisation, colours and visual interactions.

Low-level – query: The user needs to identify one or several books in the digital library. Possibly, the user can also compare books and summarize information about the library’s content.

In a case study about an energy portfolio, Brehmer, Ng, Tate, & Munzner (2016) put in practice the four-level nested model and the visualisation typology and came up with three high-level tasks that fulfil the users’ exploration needs: “overview”, “drill down” and “roll up” (p. 452). It resembles Shneiderman's (1996) three high-level tasks: “Overview: Gain an overview of the entire collection. Zoom: Zoom in on items of interest Filter: filter out uninteresting items” (p. 337). Both high-level sets of tasks are respected in the present design study because they fit the user-tasks generated when entering an exploratory search in a digital library’s metadata information space.

3.1.2.2 The data – What?
The following section aims at presenting the visible data underlying the ExploViz interface. Details about the data cleansing and data transformation processes can be found in section 3.2 The ExploViz interface implementation.

The scope of the data is first defined by the academic entry dataset related to business and finance provided by Cyberlibris. Moreover, the books’ information space is exclusively made of textual libraries’ metadata. Part of it is traditionally considered as descriptive metadata such as the documents’ titles, authors, covers and abstracts. The remaining part is traditionally considered as structural metadata such as the documents’ categories and tables of content. However, all available metadata is finally treated as syntactic data, parsed into bags-of-words.
The entry dataset can be considered as several tables of textual attributes aggregated into one table. The Figure 3.3 illustrates the number of records, unique records and missing records from the sub-sample of academic manuals collected from Cyberlibris. The Table 3.1 provides a short description about these selected fields.

![Bar chart representing the initial dataset collected from Cyberlibris, filtered on the academic manual sub-sample.](image)

### Field's name | Field's description
--- | ---
**docid** | The document’s unique identifier (0% missing)
**title** | The document’s title (0% missing)
**dicetheme** | The document’s high level theme (52% missing)
**description** | The document’s abstract (0% missing)
**toc** | The document’s table of content (12% missing)
**cover_url** | The document’s url to the Cyberlibris’ server hosting the covers (0% missing)

The original dataset used in the pre-processing steps is the same than the visible data available when interacting with the search interfaces, except for the documents’ table of content which is not shown in the final output. One should also notice that the documents’ content is not available either in the original dataset nor in the search interfaces, which limits the task and the design spaces.

Four attributes are derived from the original fields in order to enrich the bag-of-words’ semantic and improve the spatialisation. The *level 1* and the *level 2* fields are specifically used in the search interfaces’ legends. They are derived from two successive k-nearest neighbours algorithm based on the DTM and the *dicetheme*. They both form an
analytical base table where the term frequencies are the attributes and the *dicetheme* is the target variable. After the knn-algorithm is performed, a manual classification and verification is provided in order to improve the general process. The final aim is to build a two-level classification which can be used in both interfaces. Additionally, the *concept* and the *calais* fields both derive from the title, the abstract and the table of content attributes which are used as input for generating meaningful tags and keywords from two well-established web-based APIs, Aylien⁵ and Reuters OpenCalais⁶.

<table>
<thead>
<tr>
<th>Field’s name</th>
<th>Field’s description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>level 1</strong></td>
<td>The document’s high level category (manually added – 3 levels) – Derived from the <em>dicetheme</em> using 2 successive knn algorithms with ( k = 5 ).</td>
</tr>
<tr>
<td><strong>level 2</strong></td>
<td>The document’s medium level category (manually added – 9 levels) – Derived from the <em>dicetheme</em> using 2 successive knn algorithms with ( k = 5 ).</td>
</tr>
<tr>
<td><strong>concept</strong></td>
<td>Keywords describing each document, generated from Aylien OpenCalais</td>
</tr>
<tr>
<td><strong>calais</strong></td>
<td>Keywords describing each document, generated from Reuters OpenCalais</td>
</tr>
</tbody>
</table>

*Table 3.2 Derived fields from the original fields.*

The Figure 3.4 shows the transformation applied on the *dicetheme* of which 52% was initially missing.

![Figure 3.4](image)

*Figure 3.4 Transformation of the dicetheme using knn with \( k=5 \) and manual changes into the levels 1 and 2. The generated fields are used for classifying the books in both search interfaces.*

The first part of the data analysis is used to clean, filter and enrich the dataset with meaningful words. The rationale underlying the semantic enrichment process is that the number of words decreases - replacing the abstracts and the tables of content with a few meaningful keywords. As explained in the following section 3.1.3, the documents’ two

---

⁵ Aylien text classification API used for semantic enrichment: [http://aylien.com/classification/](http://aylien.com/classification/)

dimensional coordinates used to form the interactive scatterplot, result from a topic
modelling algorithm and through dimensionality reduction.

3.1.3 Encoding and interaction technique design – How?
The encoding of a visual interface is central in information visualisation because this is
what defines the visual mapping of the underlying data. The encoding defines how the
data is visually expressed and it is often subjected to design studies in the visualisation
literature.

This section first introduces the reader to the spatialisation which can be seen as the
visualisation’s ultimate substrate where distance between two documents corresponds
to their similarity. Then it highlights and justifies the encoding consisting in marks,
channels and interactivity.

As can be seen in the Figure 3.5, the suggested and evaluated visualisation is a
spatialisation representing a business academic digital library containing a sample of
1363 documents which are mainly books. The spatialisation results in a scatterplot where
each book is represented as a point. The closer two books are, the more similar they are.
In addition, three blue convex-hulls are displayed in the background. Each colour, in
combination with one convex hull, corresponds to one high-level topic. The three high-
level topics found in the studied sample are “Business”, “Finance” and “Economics”.

Figure 3.5 The ExploViz visual search interface in un-zoomed mode. This is the suggested visualisation specifically supporting exploratory search in a limited digital library. Each colour combined with a convex-hull represents a topic and each circle represents a digital document. The search bar allows the user to enter a keyword for highlighting the corresponding documents. Interactions with the interface are available such as zooming, clicking on the legend, hovering the circles, dragging the space.
3.1.3.1 Spatialisation

Two main transformations are applied on the original dataset. First, the Latent Dirichlet Allocation (LDA) algorithm generates topics from the entry DTM. Details about the LDA implementation is provided in section 3.2. The multidimensional scaling is used as means of dimensionality reduction in order to map each book with Euclidean coordinates. The books’ position and the topics’ position are relative to each other. For example, the “Actuarial science” book, highlighted in Figure 3.6, is a book related to risk. This is the reason why it is part of the Finance high-level classification (dark blue) and Corporate finance – risk low-level classification (blue sky). However, the book contains a lot of mathematical concepts. This information is integrated into the LDA process because of words such as “mathematical”, “models”, “probability”, “statistics” which can be found inside the book’s abstract. This is the reason why the book is also located near the mathematical economics region (dark green), which is part of the Economics high-level classification.

3.1.3.2 Marks and channels

The mark is the dimensional attribute of encoding, which forms the structure of the visualisation. It can either have zero (point), one (line), two (space) or three (volume) dimensions. The channel is the visual attribute which, in combination with the mark, allows for encoding the data values visually, such as the colour, the shape, the size, the angle, the position. The effectiveness of the channels was studied and measured in order
to provide a ranking which can support the designer when choosing the visual encodings (T. Munzner & Maguire, 2014, p. 95).

The main transformation applied to the original data consists in defining coordinates to each book based on their relative syntactic similarity. The points forming the scatterplot are therefore the main mark that is used in the visual interface. Besides the scatterplot’ point cloud, three convex-hulls are displayed in the background. The three areas can be considered as a mark which encodes the high-level classification. Additionally, three distinct channels are used in the visualisation. The vertical and horizontal positions define each dot’s location on the scatterplot. The axes are not displayed on the screen because the values are not meaningful individually. The spatial position is intended to form spatial regions where each book’s location is relative to the others. Additionally, the colour hue is used to strengthen the books’ affiliation to one topic. The un-zoomed mode displays three distinct colours which represent three aggregated themes. The zoomed mode displays eight distinct colours which represent the eight sub-level themes. The three convex-hulls have the same role, except they do not leverage the colour but rather the spatial region channel. The three channels used, namely the position on common scale, the spatial region and the colour hue are the three most effective channels, as reviewed in (T. Munzner & Maguire, 2014, p. 102).

3.1.3.3 Interactivity
The interactivity is also considered in the encoding and it is declined as manipulate (change, select, navigate), facet (juxtapose, partition, superimpose) and reduce (filter, aggregate, embed) by Munzner & Maguire (2014, p. 102, 242, 264, 298, 322). As highlighted in both following sections, the ExploViz interface uses the manipulate – select and navigate, as well as the reduce – embed, interactivity features.

3.1.3.3.1 Manipulate – select and navigate
The ExploViz spatialisation allows users to select points inside the scatterplot. The selection tool is one of the final outcomes of the users’ search process. Indeed, as often in a digital library, the search process ends with one or several books’ selections. The selection tool is intuitively configured by clicking on the point which represents the reader’s book of interest.

The navigation across the library is supported by a semantic zoom configured within the point cloud. The zoom allows the user to increase the scale of the visualisation from 1
to 100. The unitary scale corresponds to the overview, the scale of 100 corresponds to a detailed zoom which allows the user to distinguish books that might be overlapping in overview mode. The semantic zoom is not a magnification of the points’ size but rather a translation of their coordinates. The more the visualisation scale is increased, the more the points are translated as a multiplication factor of the scale parameter which is controlled by the user. This is the reason why two overlapping points in the overview mode can be separated apart in a detailed view. This is illustrated in Figure 3.7. Additionally, a geometric zoom is configured within the convex-hulls so that their shape’s size is magnified at the same pace than the points are translated. That way, the point cloud expands with the zoom, while always remaining inside their corresponding area.

Figure 3.7 Illustration of the semantic zoom configured in the ExploViz.

Moreover, the ExploViz is a scale dependent spatialisation. Indeed, the visual encoding changes dynamically based on the scale of the visualisation, which is controlled by the users’ zoom. When the visualisation’s zoom reaches 150%, the classification encoding changes. As can be seen in Figure 3.5, the high-level classification is encoded in the overview mode which extends from a scale of 100% to 150%. When the scale exceeds 150%, the low-level classification is encoded instead, as can be seen in Figure 3.6. The legend, which is situated in the upper right corner adapts automatically. When the encoding changes, the convex-hulls remain unchanged in order to keep track of the high-level classification’s area. It is a way to prevent the user from getting lost inside the information space, as landmarks would.

3.1.3.3.2 Reduce - embed
A point cloud alone in combination to a legend does not support the users in getting individual books’ information. This is why an embedded card containing the books’ title, author and abstract is configured within the point cloud on mouse hover. As can be seen
in Figure 3.6, the card appears when a point is hovered with the mouse. The card’s colour corresponds to either the high or the low level classifications depending on the visualisation’s scale, for strengthening the book’s classification. As soon as the mouse is away from the point, the card disappears.

Additionally, it was decided to add a keyword search bar combined to a “search books” button on top of the visualisation’s frame (Figure 3.8) in order to let the users focus on keyword-based queries’ results. The bar is the same than any traditional web search engine or digital library search entry point. The keyword search tool is configured so that the books resulting from the user’s query are highlighted inside the visualisation instead of filtered away. That way, the resulting search does not modify the user’s working context. Instead, it reduces the opacity of the points which are outside the search results in order to make the books of interest more visible. This process differs from the traditional use of a search bar which only provides the resulting space of results. It can be seen as a filter tool which does not remove the working context. First, a search bar was not considered as it resembles the list-view baseline core process. Finally, it was implemented as a tool that supports navigation across the library’s space while protecting the context. It can be seen as a focus + context tool rather than a traditional filter tool.
The Figure 3.9 is an illustration of both four-level nested model and the visualisation multi-level typology adapted to the ExploViz search interface. It shows a “chained sequence of what – why – how analysis” inspired from the Glimmer visualisation tool which mainly aims at discovering clusters of similar documents (T. Munzner & Maguire, 2014, p. 318). The ExploViz interface also allows for navigation in the library’s metadata space, contrary to Glimmer which focuses on the resulting clusters from dimensionality reduction. This sequence of what – why – how can be considered as a design justification for ExploViz in the form of a design study. The first sequence describes the data transformations needed to generate the ExploViz entry dataset. The second sequence resumes what – why – how that were discussed above. Following the ExploViz interface design, the next section presents the ExploViz interface implementation.

3.2 The ExploViz interface implementation

The following section introduces the reader to the practical implementation of ExploViz, from the data collected into R, to the visual interface’s development in D3. First some
details about the data collection are provided. Then, the data profiling followed by the
data cleansing and semantic enrichment processes are explained. Next, the data
transformation based on topics’ modelling and dimensionality reduction are presented.
Finally, the ExploViz graphical user interface’s implementation in D3 is exposed in a
 nutshell.

3.2.1 Data collection
The data input used to set up the experiment is provided by Cyberlibris. The raw dataset
consists of 12 csv files, web-based cover images and xml tables of contents. The
complete dataset tallies approximately 8000 business related documents and 80000
users. Information about the documents as well as anonymised information about the
readers can be found in the original dataset. However, it is decided to filter the input data
for the project in order to work with the most consistent data for building the most
relevant experiment.

3.2.2 Data profiling and sample selection from the original dataset
It is decided that only the books’ information from the original dataset will be used in
the following experiment. Indeed, using Cyberlibris reading behaviour log data would
add unnecessary complexity, especially since traditional digital libraries do not usually
leverage the readers’ information. Moreover, only the manuals are kept among other
document types. Indeed, the manuals are considered to be the most representative digital
documents in an academic library and they are easier to approach for non-specialist users
compared to scientific articles. Finally, the records containing missing title, missing
abstract or missing table of content are removed from the dataset. After a complete data
cleaning, wrangling and analysis, 1363 documents form the final search interface input
dataset.

3.2.3 Data cleansing and data standardisation
As expected, the dataset pre-processing task had been a long and meticulous work as it
usually is in data analytics projects. All data cleaning, data wrangling and data analysis
is made in R, using the R Studio interface7. It is decided to use mainly textual information

7 R Studio: https://www.rstudio.com/
for spatializing each document on a 2D plane. For tracking purposes, the R files are available on GitHub online\(^8\).

3.2.3.1 Data import in R
Import the books’ related information. The entry files are CSV with a semi-comma separation. The encoding is UTF-8 and it is imported as such in R. Three flat files are imported: the books.csv file, the doctype.csv file and the abstract.csv file. Additionally, the tables of content are requested online via an R loop. The four textual information related to documents are stored in distinct R data frames.

3.2.3.2 Text cleansing
The textual data cleaning is an essential task as later processes will use bags of words as input. Therefore, the cleaner the bag of words, the more accurate the semantic outputs.

- First, the punctuation and the numbers are removed from each string using the R “stringr” package in combination with regular expressions ("^[[:alpha:][:blank:]]")
- Then, the one-to-three letter words are removed using a regular expression ("^[[:alnum:]][1,3][[:alnum:]*]"
- Letters are transformed to lower case.
- English stop words are removed using the R “tm” package
- White spaces are removed

3.2.3.3 Dataset semantic enrichment
A clean textual dataframe is created assembling the docid, the title, the abstract and the table of content of each manual. It is called bow as bag-of-words. A new field is created, combining the three information. It is called “semantic.input” and it is used as input parameter for two text analysis APIs. Two loops are developed in R for extracting semantic keywords from open source API and linking them to each corresponding document. Two distinct APIs are used in order to diversify the outputs. The concepts extraction from the Aylien text analysis API and the social tags extraction from the Reuters Open Calais API. The latter consists in matching the Wikipedia folksonomy, which is the Wikipedia’s collaborative tagging system\(^9\). The semantically meaningful

---

\(^8\) R files used for the ExploViz project cleaning and transformation: [https://github.com/klem88/ExploViz_Dissertation](https://github.com/klem88/ExploViz_Dissertation)

tags are associated to each document based on their resemblance with Wikipedia’s articles. Both applications are fed with the same information return semantic metadata which are used as input for the spatialisation process. The two semantic results are stored in the bow dataframe. High level descriptions like topics or industries were also extracted using the title, the abstract and the table of content information but the results were either highly missing or not relevant. On the contrary the social tags and concept, which are lower level descriptions, are not missing and show accurate responses. The latter are therefore chosen to be the next step’s input.

The previously calculated concepts and social tags are combined with the title in a unique field called bow1. The strings are transformed to lower case and the English stop words are removed in order to keep only meaningful words.

In addition to the titles and the semantic keywords, two fields called “dicetheme” and “subject” from the original dataset assign one high level theme and one low level subject to the library’s books. Unfortunately, this information is highly missing (52% of missing themes and subjects). Yet, it is a good opportunity to consider high and low level clusters. It is therefore decided to extract the dataset and to manually label clearly identified documents with the chosen two-level classification from the available themes and subjects. This operation is also a good opportunity to delete the books that are irrelevant with the chosen clusters.

The manually modified document is imported back into R and a k-nearest neighbour algorithm is applied on the remaining missing values using the “class” R package. The results are then manually checked and modified where needed. The process is repeated three times in order to refine the dataset. The high level classification is called “Level1” and the sub-level classification is called “Level2”.

57
3.2.4 Data transformation – Data spatialisation

3.2.4.1 The document term matrix

![Data Quality Chart - Entry DataSet for LDA](image)

Figure 3.10 Data quality chart displaying the entry dataset for LDA

All following transformations are based on a document term matrix (DTM) which breaks the document corpus down into the words composing all documents. Each cell within the DTM matrix represents the word-count of the corresponding term and document. It is assumed that the most representative words of each book are contained into the title, the level1 and level2 fields. Additionally, the concepts and the social tags attributes should also be representative of their corresponding book’s abstract and table of content. Therefore, the fields level1, level2 and bow1 (title, concepts and social tags) are entered inside the DTM. The five fields statistics are illustrated in Figure 3.10. The DTM is constructed using the R “tm” package. After removal of English stopwords, it is a 100% sparse matrix containing 5063 terms and 1363 documents, as illustrated in Figure 3.11. The least representative terms are removed in order to reduce the sparsity of the matrix and to reduce the following calculation’s complexity. With a sparsity of 95%, the matrix contains 205 terms which still represent all the documents, meaning that each document at least contains one of the remaining 205 terms.

```
<<DocumentTermMatrix (documents: 1363, terms: 205)>>
Non-/sparse entries: 15140/264275
Sparsity : 95%
Maximal term length: 14
weighting : term frequency (tf)
```

Figure 3.11 Sparse document-term matrix’s (sparse.dtm) characteristics accessed in R (02/08/2016)
The following process aims at spatializing the books on a two dimensional plane and classifying each book into one high-level theme and one sub-level theme for navigation purposes. In other words, each book’s 205 attributes need to be transformed into a two dimensional coordinate (x, y) and need to inform a hierarchical two-level thematic structure. Before choosing the best matching transformation, several trials are performed. First the DTM is used, without intermediary, as the source for clustering algorithms (k-means, walktrap community) and distance measurements (cosine, Euclidean, Jaccard). Different weights have also been applied on the DTM (TF, IDF, TF-IDF, log(TF)-IDF) in order to test different weighting schemes. The resulting visual spaces were either too aggregated or overly homogeneous with difficulties to make out academic classifications. This is why it was decided to add a preliminary step before identifying themes and computing distances between the books. Probabilistic topic models offer a way to extract topics from an underlying information space, namely the DTM, and to map each document with probabilities to approach each topic.

3.2.4.2 The Latent Dirichlet Allocation (LDA) topic modelling.

The Latent Dirichlet Allocation (LDA) is probably the easiest and the most documented topic modelling technique. It is therefore chosen for classifying and spatializing the documents. The number of generated topics needs to be provided as input to the LDA algorithm. Grimaldi, Corvello, Mauro, & Scormozzino (2015) elaborated a simple strategy for identifying a two-level theme structure out of the topics’ probability matrix generated from the LDA algorithm. The same methodology is followed. It consists in computing the Pearson’s correlation coefficient between each pair of topics’ probabilities associated to each document. The symmetrical correlation matrix is transformed into a dissimilarity matrix by subtracting the correlation values from one. Finally, the R hierarchical clustering algorithm is performed and plotted on the dissimilarity matrix in order to visualise affinities between the themes. The correlation matrix shows how much pairs of topics fluctuate together. A positive correlation between two topics indicates that their probabilities increase together. Moreover, if a document has high chances to be related to one of both topics, it is also likely that it has high chances to be related to the other one. Positively correlated topics are likely to

---

10 A short description of the initial data transformations and a selection of resulting spaces are showed in appendices 7.2 Resulting visual spaces of Document-Term Matrix initial transformations.
appear together in the same document. Consequently, this is why it is assumed they are likely to be semantically closer.

![Hierarchical clustering](image)

*Figure 3.12 Hierarchical clustering on the dissimilarity correlation matrix.*

Based on this assumption, and based on the visual clustering (Figure 3.12), each theme is manually labelled based on levels 1 and 2 distributions among the nine topics resulting from LDA and based on the author’s judgement. The Table 3.3 illustrates the combination between the LDA topics and the three level 1 labels assigned in the entry dataset. As can be seen, the 9 clusters created by the LDA algorithm match the three level 1 labels.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Business</th>
<th>Economics</th>
<th>Finance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>187</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>56</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>198</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>110</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>156</td>
</tr>
<tr>
<td>6</td>
<td>181</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>128</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>3</td>
<td>113</td>
</tr>
<tr>
<td>9</td>
<td>172</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

*Table 3.3 Level 1 distributions among the nine topics resulting from the LDA performed on the DTM*

It is decided that the resulting output is acceptable both in terms of visual spatialisation and in terms of thematic classification. It was finally accepted after several tests on both
the number of topics and the DTM’s sparsity. During the tests, the number of topics varied from 3 to 20 topics and the DTM’s sparsity varied from 90% to 100%. Finally, 95% sparsity and 9 topics, manually reduced down to 8 topics, provide an acceptable match for the requested visual thematic spatialisation needed to perform the experiment.

In order to be consistent, the spatialisation operation is based on the previously generated LDA’s output. First the Euclidean similarity matrix is computed from the topics’ probability matrix. Then a multidimensional scaling algorithm (cmdscale in R) is applied as a means of dimensionality reduction from the Euclidean similarity matrix to a two dimensional plane. The resulting dataset therefore corresponds to the list of underlying documents linked to two dimensional coordinates x and y.

3.2.5 The ExploViz graphical user interface

The ExploViz graphical user interface is fully built using the Data Driven Document (D3) language. The D3.js\textsuperscript{11} is a JavaScript graphical library optimised for creating and displaying Scalable Vector Graphics (SVG) based on datasets. The ExploViz script is based on two datasets. The first one corresponds to the list of documents linked to two dimensional coordinates and metadata such as the title, abstract and themes. The second one corresponds to the three convex-hulls two dimensional coordinates.

The implementation can be divided in three parts. First the convex-hulls defining the high-levels documents’ categories are displayed on the background. Then, each record from the documents’ dataset is spatialized as a point on the plane, encoded within the category’s corresponding colour. Finally, the legend is printed on the top right of the page. Additionally, the interaction features are easily implemented within D3 using the JavaScript mouse events on the points and the legend. Moreover, D3 offers a built-in zoom option which is used for implementing the semantic zoom.

The successive versions of the ExploViz interface code scripts and datasets are stored and available on GitHub. The version 16 out of 17 is the version that is used in the experiment\textsuperscript{12}.

\textsuperscript{11} Data Driven Document: https://d3js.org/ (accessed on 11/08/2016)
\textsuperscript{12} ExploViz version used in the experiment: https://github.com/klem88/Visual-Digital-Library/tree/version-16 (accessed on 11/08/2016)
3.3 Summary

This chapter first justified the ExploViz interface design using the *four-level nested model* framework. The visual search interface aims at supporting the users’ exploratory search behaviour when interacting with a digital library’s content. The user-tasks are described as *discovery* and *enjoyment* when exploring the underlying information space. The documents’ metadata and location are encoded within the point cloud whereas the topics are encoded within the colours and convex-hulls. Additionally, *selection* and *navigation* features are implemented, as well as a semantic zoom and focus + context options such as the card on demand and the search bar which allow for interaction within the underlying collection of documents. Secondly, the chapter explained the ExploViz interface implementation from the entry dataset collection to the implementation of the graphical user interface. Topic modelling is used in the ExploViz interface development as an intermediary tool for extracting hierarchical topics from the collection of documents and for defining relative document similarity by assigning them 2D coordinates. Following the ExploViz interface design and implementation, the next chapter 4 addresses the ExploViz interface evaluation and more generally the project’s hypotheses evaluation.
4 EVALUATION

The evaluation of exploratory search systems remains a challenge, as first identified at the ACM SIGIR forum of 2006. In this workshop, a model for exploratory search process was considered coupled with possible metrics’ such as “Engagement and enjoyment”, “Information novelty”, “Task success”, “Task time” and “Learning and cognition”. Part of the following section reflects the guidelines developed during this workshop (White, Muresan, et al., 2006, p. 58). Moreover, Munzner (2008, 2009) outlines the possible threats of missing evaluation at different levels of her four-level nested model in information visualisation papers. She provides upstream as well as downstream validation guidance at each level, which is followed in the evaluation chapter.

This chapter first section 4.1 discusses the experimental design which aims at testing the impact of visualisation on the process of exploratory search. The experiment’s configuration is outlined and the three user-tasks, as well as the cognitive load measurement are described and justified. Finally, the five evaluation metrics are considered in depth. As a comparison, the section 4.2 introduces the design and the implementation of the baseline LibSearch interface. Next, the section 4.3 focuses on the experiment’s implementation and the section 4.4 presents and discusses the results. Finally, the section 4.5 concludes the chapter.

4.1 The experimental design

The experimental design aims at testing the potential influence of spatialisation on exploratory search. This is done by comparing the users’ behaviour when interacting with two distinct search interfaces - one visual interface and one baseline list-view interface. It is therefore decided to perform a between-group comparison based on a set of exploratory search tasks. The potential difference is measured using a set of novel quantitative metrics and further examined with a set of well-established cognitive load metrics.

The following section first presents experimental approach. Then, the three user-tasks are described and justified with respect to the nested-model and multi-level typology. Finally, the cognitive load measurement system and the evaluation metrics are discussed.
4.1.1 A between group experiment comparing two search interfaces

It is common in the literature to compare an application with its baseline corresponding application. For example, Y. Liu et al., (2013) and Ruotsalo et al. (2013) performed a formal evaluation of their respective visual search interfaces in comparison to non-visual baselines. This type of comparison is not usual in the literature as visualisation design studies tend to evaluate the efficacy of the design with other visual tools. Quantitative indicators are used across each interface to ensure consistency and enable statistical conclusions. Both search interfaces are based on the same dataset, but the information is presented to the user differently. The list view search interface displays the books’ information as a list of books, whereas the spatialized search interface displays the books on a two dimensional plane.

For a long time, researchers have focused on exploratory tasks’ development rather than their evaluation (White, Marchionini, & Muresan, 2008, p. 433). Traditional metrics, such as precision and recall, are not particularly useful when assessing exploratory search behaviours because exploratory search does not result in a correct or incorrect answer. On the contrary, a successful exploratory search might include periods of discovery, serendipity and sensemaking. However, even though exploration is difficult to assess in absolute terms, exploration can be compared across applications to assess which supports or encourages exploration. Those are the reasons why it is decided to compare the ExploViz visual search interface with a non-visual list-view search interface.

A between group configuration is chosen in order to avoid the learning effect where the users learn the tasks’ answers while using the first interface in the experiment. Although counterbalancing is a possible solution to this bias, the experiment would long and repetitive to perform (MacKenzie, 2012, p. 177). As a result, only one interface is presented to each group. Finally, the user is unable to select a given interface to avoid the novelty effect whereby the user would choose the new and original interface (Jha, 2014, p. 155).

4.1.2 An online task-based experiment

The experiment’s online setup offers the possibility to perform the tasks in each user’s real life context. This is assumed to provide a more realistic experience for the user. Even though the online setup can reduce control, it is believed that the benefits of a
realistic framework outperform the disadvantages of weaker experimental control. The experiment’s online setup also offers the possibility to reach a larger sample of respondents, who are more readily accessible. Finally, this approach could potentially decrease chances of the *Hawthorne effect*, where the respondent’s behaviour is biased knowing s/he is observed (Reips, 2000, p. 92).

4.1.3 The user-tasks

The user-tasks are a combination of visualisation and information-seeking tasks. They are similar to the previous model’s task abstraction, except that each task needs to be achievable and workable in both search interfaces; one visualisation and one list-view. The Figure 4.1 outlines the impact of the *multi-level typology* framework on the user-tasks designed for the between-group experiment using both a visual and a list-view search interface. This is a significant constraint because the visualisation tasks such as defining clusters, zooming or find outliers cannot be performed within the baseline list-view. The experiment requires identical tasks that can be performed comparably in both interfaces. Given that an aim of this thesis is to approach the problem using a quantitative methodology, the ExploViz experiment prioritises structured input from the users ensuring a comparable. Therefore, no qualitative comment is retrieved during the experiment.

![Figure 4.1 The experiment's user-tasks detailed through the multi-level task typology](image-url)
The three tasks are presented below. The first task consists of a sensemaking task, as adapted from Y. Liu et al.’s sensemaking task (2013, p. 29), and does not require an input search-query from the respondents. The second task is an open-ended and complex directive scenario-based task, with no keyword given as a hint inside the instructions. Finally, the third task is a book selection task, which was designed to reflect the definition of exploratory search (Ryen W. White & Resa A. Roth, 2009, p. 10).

4.1.3.1 Sensemaking task 1
As presented previously, sensemaking is close to exploration. It is assumed to be similar to exploratory search in this context. The user is asked to find representative keywords for the three unlabelled topics available in the digital library. It does not ask the user to enter a search specific to one document but rather asks the respondent to develop an appreciation for the information space, exploring the documents’ metadata, a common component of exploratory search. Figure 4.1 details the sensemaking task using both the nested model and the typology terms. Discovery is used as generating hypotheses about each topic’s most representative keywords.

The search interface is a digital library containing 1363 academic books. The library can be broken up into 3 high-level topics. Exploring the library, find 5 keywords that best describe each topic. You can use the representative words from the titles and the abstracts.

4.1.3.2 Directive scenario-based task 2
Documents’ content is often available and accessible from visual search tools. Therefore, user-tasks often require to enter the documents in order to find the answer. For example the information gathering task presented by Y. Liu et al. (2013, p. 29) and the specific question task suggested by S. Liu et al. (2012, p. 25) requires from the users to look inside the digital documents, whether articles or emails. Other user-tasks such as the directive scenario-based tasks designed by Hoeber & Khazaei (2015, p. 236) provide the respondents with a context and an input keyword and ask them to find relevant papers. With ExploViz, specific information gathering cannot be asked because the documents’ content is not provided. However, a scenario-based task is possible as long as the requested information can be found in the books’ titles and abstracts.
The second task asks the user to find two books given a specific scenario. The task is open-ended and does not provide specific details nor a suggested entry keyword as in Hoeber & Khazaei (2015, p. 236). It can also be interpreted differently because it relates to two possible topics, namely *team management* and *financial instrument* without specifying any further restriction. As outlined in Figure 4.1, the low-level task is about identifying digital documents rather than summarising information. Users are expected to use the keyword filter tool and the similar books information.

“You are a financial auditor and you have just been hired as a team manager in a multinational bank dealing with financial instruments. This is the first time that you will work in a bank and that you will manage a team of financial experts. Exploring the library, select books that would help you understanding your new role.”

4.1.3.3 Free exploratory search task 3

The free exploratory search task is designed so that respondents express their search behaviour. The task consists of selecting one book from the digital library and an e-book is offered to the random draw’s winner. The raffle is used in order to motivate the users to engage in a more comprehensive search strategy. Contrary to the other tasks, this one is based on the respondent’s enjoyment as outlined in Figure 4.1.

“Exploring the library, choose one and only one book that you would like to read. If you enter the raffle, you might win the e-version of the book you choose.”

4.1.4 The cognitive load measurement

In addition to specific tasks, Exploratory visualisation systems (EVS) can be evaluated and compared using elements borrowed from cognitive load theory (Sweller, 1988). Cognitive load is an ill-defined concept (Longo, 2014) and its representation, measurement and inference are not trivial (Longo, 2015). However, it is a widely invoked design concept and it has been employed for various purposes, including enhancement of usability methods (Longo & Dondio, 2015) and evaluation of web-interfaces (Longo, Rusconi, Noce, & Barrett, 2012). These example are strictly connected to the work presented in this thesis. In this instance, in particular, after each task is completed, some of the questions of the well-known NASA task load index (TLX), developed by Hart & Staveland (1988), are used to establish the cognitive load that the task put on the user. It consists in asking a set of 5 points Likert scale questions about the user’s state when performing the task. The results aim at providing levels of
intrinsic, extraneous and germane cognitive loads, which explain the user’s mental load allocation while performing a task. The evaluation of an EVS especially addresses both the extraneous and germane loads. The extraneous cognitive load explains the mental load allocated to understanding and interacting with the search interface. A good EVS therefore shows low levels of extraneous cognitive load. On the contrary, the germane cognitive load explains the mental load allocated to analysing information and learning new knowledge. A good EVS therefore shows high levels of germane cognitive load (Y. Liu et al., 2013).

Two performance questions, measuring the germane cognitive load, added to the effort and the frustration level questions, measuring the extraneous cognitive load are presented to the respondents after each task completion. Comparing the resulting cognitive loads of both search interfaces enables us to assess the cognitive load that each interface placed on users. Contrary to Y. Liu et al. (2013, p. 31), the mental and physical demand questions are not asked because they do not allow a clear distinction between extraneous and germane cognitive loads if not associated with some users’ personal comments. Additionally, the temporal demand was removed during the pre-evaluation phase when the experiment was tested with a very small sample of respondents, following which it was noticed that no time constraint applies on the tasks. The tester’s feedback about the temporal question was the following: “The persons who runs the test also defines the pace. Time pressure does not come from the test but from the respondent who defines how much time to allocate to the given task.”

4.1.5 The evaluation metrics
4.1.5.1 Measuring exploratory behaviour
The respondents’ hover events are stored at each task from 500 milliseconds. The project’s hypothesis assumes that the use of spatialisation while performing exploratory search tasks increases the likelihood of interacting with the interface through hover events, exploring more of the information space and engaging with more of the library’s content. Two metrics are derived from the respondents’ hover events. The number of hovered items as well as the underlying area explored defined by the hovered items are persisted while the tasks are performed. The assumption is that the mouse-hover action is equivalent between both interfaces. The LibSearch space which triggers the hover event corresponds to the books’ titles, covers and abstracts displayed in the list-view,
whereas the ExploViz space which triggers the hover event corresponds to the books’ points. Both are deemed equivalent.

4.1.5.1.1 The number of hover events
The number of hover events is measured and compared between the two interfaces in each task. All the stored events are used in the computation from 500 milliseconds and above, per interface and per task.

4.1.5.1.2 The Area explored

![Illustration of the area explored metrics and its equivalency between the list-based and the visual search interfaces](image)

The area explored is defined as the surface covered by the convex-hull surrounding the point cloud representing the books that are processed by the users while performing the task. Figure 4.2 illustrates the definition of the area explored. The books’ ids are recorded with each hover event. Indeed, it is assumed that all hovered books take part in the respondents’ process for answering the questions. The area explored is computed and compared between groups in each task. Additionally, the area explored is measured for each respondent and each topic separately in order to avoid counting for empty exploration space such as the situation illustrated in Figure 4.3.
4.1.5.2 Syntactic and semantic proxies for sensemaking accuracy

The respondents’ answers, considered as bags-of-words, are analysed both syntactically and semantically. The syntactic proxy focuses on the words’ representativeness based on the topical term frequencies, whereas the semantic proxy focuses on the meaning of the given words.

4.1.5.2.1 Syntactic sensemaking proxy

The users’ answers are gathered as a bags-of-words. Each answer’s bag-of-words is compared to its related topic’s bag-of-words in order to measure how representative the respondent’s answer is. The answers’ accuracy is therefore measured as the bags-of-words’ representativeness against its corresponding topic. Both tools’ accuracies are then compared.

The traditional metrics such as the Jaccard index or the Euclidean distance are not well adapted for expressing the representativeness of bags-of-5-words against a much larger corpus. Computing the Jaccard index would effectively involve counting the number of words co-occurring between the topic and the answer. The latter would generally be 5 and would not measure the representativeness of the words. The Euclidean distance,
would measure the difference between two completely different words’ distributions without capturing the right information. Therefore, the computed index is decided to be the sum of the topic’s term frequencies co-occurring between one answer and its related topic. The Figure 4.4 illustrates the essence of the computed syntactic metric. As can be seen, the most representative respondent’s bag-of-words is the one that gathers words that have the biggest term frequency in the corresponding topic’s bag-of-words. The stemming and English stopwords transformations are equally applied on the bags-of-words.

![Figure 4.4 The figure illustrates the essence of the syntactic index measuring the topic’s representativeness of small sets of words.](image)

4.1.5.2.2 Semantic sensemaking proxy

Three distinct types of semantic proxy are designed. The first one uses topic modelling (LDA) and follows the same process as used when developing the spatial representation of the ExploViz library. The process consists in computing the probabilities of each respondent’s bag-of-words toward the underlying sub-topics. The number of underlying topics is set up to five so that each respondent can potentially use one word for each topic. The same document-term-matrix as illustrated in Figure 4.4 is used as input for the LDA Gibbs algorithm. The Euclidean distance matrix is then computed based on the resulting probability matrix. The distances from each respondent’s bag-of-words to the corresponding topic’s bag-of-words are statistically compared between both groups. The operation is repeated for each topic.
The second semantic sensemaking proxy is based on the latent semantic analysis (LSA) algorithm. The algorithm takes the same document-term-matrix as input (Figure 4.4). The cosine similarity metric is computed on the low dimensional space resulting from the LSA algorithm. The similarity measures of each respondent’s answer against the corresponding topic’s bag-of-words are statistically compared between both groups. The operation is repeated for each topic.

Finally, the semantic text similarity measure provided by the Dandelion API\textsuperscript{13} is also used. The underlying process uses an ontologically driven semantic service to define the semantic similarity between two bags-of-words. The topic’s bag-of-words is reduced in order to comply with the maximum query limitation of 4016 characters. After removing the stopwords, the document-term-matrix is sorted in descending order of term frequencies so that each topic’s most frequent words are part of its representative bag-of-words. The API is then queried in order to get the similarity measure for analysis. Both groups are then statistically compared.

4.1.5.3 The distance between two selected books

In the second scenario-based task, the respondents are asked to select two books in the library. The distance between the two selected books is computed as the Euclidean distance between the two books’ coordinates. The project’s hypothesis assumes that the use of a spatialisation diagrams while performing exploratory search increases the likelihood of selecting farer apart books.

4.1.5.4 The exploration time

In the context of exploratory search, the project assumes that exploration time is an indicator of the search interface’s support for exploratory behaviour. In other words, the longer the exploration time, the better support provided while performing the given free exploratory search task. The time spent for completing the third task is therefore recorded and compared between both groups. Indeed, the amount of time is considered as exploration time providing information about the users’ curiosity and motivation while looking for a book inside the digital library.

\textsuperscript{13} The Dandelion Text Similarity API: https://dandelion.eu/docs/api/datatxt/sim/v1/ (accessed the 20/08/2016)
4.2 The baseline search interface: LibSearch

Figure 4.5 The baseline list-view search interface home page. (a) The search interface provided in the sensemaking task. The user does not have access to the topics’ labels but can still distinguish the topics’ clusters. (b) The search interface provided in the second and third tasks. The legend is complete and informs the users about the books’ classification.

For comparison purposes, a list-view equivalent search interface was developed. The ExploViz’ spatialisation provides information about each book’s similar books and classification. Therefore, the same feature was developed in the LibSearch list-view interface. A recommended books section was developed in order to mimic the similar books’ information given through the books’ positioning, as illustrated in Figure 4.6. Moreover, as can be seen in Figure 4.5 (b), a dropdown list showing the books’ classification was implemented in order to provide the same information than the ExploViz’ convex-hulls. Moreover, as illustrated in Figure 4.5 (a), the dropdown list is modified in the sensemaking task 1 in order to hide the categories’ labels.
The baseline search interface is designed to be equivalent to the ExploViz interface, but without the inclusion of the spatialisation visualisation. The ExploViz shows each book’s classification, as well as each book’s similar books. This is the reason why the dropdown list was designed and added to the list-view search interface, so that each topic can be searched separately. ExploViz provides the users with a second level classification, which is not included in LibSearch. Indeed, its inclusion would have caused some confusion. Each book’s similar books can be found in the LibSearch when clicking on the book of interest. It is also possible to click on a similar book to access its details and its similar books. The ExploViz interface also shows the books’ covers, titles, authors and abstracts on demand (when the points are mouse hovered). Equivalently the LibSearch interface displays the books’ title, cover and abstract. When clicked, the details are fully displayed, as well as the similar books’ section and a button that allows the user selecting the book (Figure 4.6).
4.3 The experiment implementation

The experiment is implemented online. First a consent page presents the experiment’s general information as well as the consent policy that need to be agreed by the potential respondents before starting the experiment. At this point, there is one in two chance of reaching the LibSearch interface or the ExploViz interface. If the experiment is started within the ExploViz interface, a 1-minute introductory video explaining how to use the ExploViz interface is presented to the user.

Figure 4.7 The first task visual instructions corresponding to the underlying ExploViz interface.

Each one of the three tasks found in the experiment can be broken down into three successive parts. First a large popup page is presented to the respondents, while the underlying interface is loading (Figure 4.7). It contains the corresponding task’s instructions. The latter are visually and attractively displayed in order to whet the respondents’ curiosity. The instructions’ visuals are adapted to the underlying search interface, but their meaning are the same. Then, either the LibSearch or the ExploViz interface is displayed on the full page for performing the task. Finally, the respondents click the “Begin Task Two” button, as shown in the Figure 4.8 and are directed to the four Likert scale questions measuring their cognitive load while performing the task.

14 Experiment’s URL: http://ditviz.com/exploroviz/home
15 The ExploViz 1-minute introductory video presented in the experiment: https://www.youtube.com/watch?v=EDfBYwPtJnw
Those three parts are repeated three times – once per task. Finally, the respondent is invited to leave an email address for participating to the raffle.

Figure 4.8 The first task respondent’s keyword input and “Begin Task Two” button.

4.4 Results

4.4.1 Data preparation

The data is retrieved from the server’s MySQL database using MySQL Workbench. Only the fully-completed answers’ records are extracted in order to avoid missing data. Four tables are exported in CSV format:

- **The users’ input:** the keywords’ input from task 1, two selected books from task 2 and one selected book from task 3, per respondent ID.
- **The cognitive tasks’ results:** a set of 4 Likert scale answers (from 1 to 5) per task and per respondent ID.
- **The respondents’ events:** the hover events and the hover time, per respondent ID and per task.
- **The exploration time for task 3:** a two-columns table containing the respondent_id and the exploration time for task 3.

The misspellings and the typos found in the users’ keyword input are corrected manually so that each respondent has a comparable bag-of-5-words in regards to the syntactic and semantic similarity measures. For example, the main misspelling was “ressources” while
subjects of concepts such as “maths” were translated into “mathematics”. One answer is considered an outlier and was thus removed because the words entered for the first task were completely outside the scope of the library for the keywords included food, vegan, fruit, calories, heartrate, swimming. None of the entered words is representative of the library’s underlying topics. It is therefore decided to remove this record from the results’ dataset, which contains 54 clean and complete answers. The four documents are then imported into R for data manipulation and statistical comparison between both groups.

4.4.2 Data manipulation

The respondents’ keywords are gathered in a matrix, similar to a document-term matrix filled with term-frequencies. The corresponding topic’s titles, abstracts, table of contents, semantic keywords, levels 1 and 2 classifications are included in the matrix. The English stopwords are removed and the remaining terms are stemmed, resulting in a matrix similar to the DTM in the Figure 4.9. The users’ part of the matrix is normalised to a binary value to account for respondents who repeatedly entered the same word.

The syntactic proxy for sensemaking corresponds to the product of the binary part of the matrix by the transposed topic’s term frequencies vector, as outlined in the Figure 4.9. The resulting matrix therefore provides each answer’s representativeness per respondent ID and per topic. This operation is repeated for each topic. Either a parametric or non-parametric test for comparing two independent groups is applied on the ExploViz and LibSearch distributions, depending on both distributions’ hypothesis of normality.
4.4.3 Statistics

Given the configuration of the experiment, both groups of answers can be considered as two independent variables. One respondent interacts with one interface only and has no relationship with others respondents. The methodology used for computing the results’ statistics is the following for each metric. First each distribution’s hypothesis of normality is tested in order to define whether to use a parametric or a non-parametric test for comparing distributions. If both groups’ distributions accept the hypothesis of normality, the Student’s t-test is applied on both distributions. If one or both groups’ distributions are rejecting the hypothesis of normality, the Wilcoxon rank sum W-test is applied on both distributions. All the statistical tests are interpreted based on a 5% alpha level of confidence. The records falling outside the boxplots’ lower or higher whiskers are decided to be included in the distributions because they express the respondents’ meticulousness rather than extreme or erroneous behaviours.
The following section presents each task’s statistics separately, with a discussion where appropriate. The tables presented below show both the normality and the comparison tests’ results as well as the median for non-parametric tests or mean for parametric tests. The tables’ cells are coloured depending on whether the null hypothesis is rejected or accepted. Each metric is illustrated through boxplots across the tasks (included in the appendices in section 7.1). Finally, a concluding table displays an overview of all results and a concluding discussion.

4.4.3.1 Task 1 statistics

<table>
<thead>
<tr>
<th>Task 1 statistics</th>
<th>Search</th>
<th>Viz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hover events</td>
<td>$W = 0.92, p = 0.05167$</td>
<td>$W = 0.64, p &lt; .001$</td>
</tr>
<tr>
<td>Area explored</td>
<td>$W = 0.93, p &lt; .001$</td>
<td>$W = 0.94, p = 0.001214$</td>
</tr>
<tr>
<td>Extraneous cognitive load</td>
<td>$W = 0.91, p = 0.001479$</td>
<td>$W = 0.85, p &lt; .001$</td>
</tr>
<tr>
<td>Germaine cognitive load</td>
<td>$W = 0.90, p &lt; .001$</td>
<td>$W = 0.87, p &lt; .001$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wilcoxon rank sum test</th>
<th>H0: Same distribution</th>
<th>Median - (Search, Viz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hover events</td>
<td>$W = 328.5, p = 0.5611$</td>
<td>60, 69</td>
</tr>
<tr>
<td>Area explored</td>
<td>$W = 2349, p = 0.01743$</td>
<td>0.0057, 0.0081</td>
</tr>
<tr>
<td>Extraneous cognitive load</td>
<td>$W = 1620, p = 0.2779$</td>
<td>3, 3</td>
</tr>
<tr>
<td>Germaine cognitive load</td>
<td>$W = 1139, p = 0.0431$</td>
<td>3, 4</td>
</tr>
</tbody>
</table>

Table 4.1 Task 1 statistics about the behavioural metrics as well as the germaine and extraneous cognitive loads. The tests’ results as well as the median are shown. The cells are coloured depending on the tests’ significance.

4.4.3.1.1 Level of exploratory behaviour

As outlined in Table 4.1, the respondents’ number of hover events distributions from task 1 do not show evidence of a significant difference between both ExploViz and the LibSearch interfaces. However, the area explored while performing task 1 is significantly different whether the respondents interact with the LibSearch or the ExploViz interface. Respondents interacting with the ExploViz interface explored a significantly larger area than respondents interacting with the LibSearch interface (as outlined in Table 4.1). This result corresponds to the project’s assumption whereby the bigger the area explored, the more supportive the interface towards exploratory search.
### 4.4.3.1.2 Proxies for sensemaking accuracy

Three bags-of-words – one per topic – were collected from the respondents while performing the task 1. Each respondent’s bag-of-words is compared to the corresponding topic’s bag-of-words using 4 distinct metrics – one syntactic metric and three semantic metrics.

<table>
<thead>
<tr>
<th>Syntactic</th>
<th>Search</th>
<th>Viz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>$W = 0.79, p &lt; .001$</td>
<td>$W = 0.93, p = 0.05579$</td>
</tr>
<tr>
<td>Topic 2</td>
<td>$W = 0.86, p = 0.003161$</td>
<td>$W = 0.98, p = 0.8792$</td>
</tr>
<tr>
<td>Topic 3</td>
<td>$W = 0.91, p = 0.02714$</td>
<td>$W = 0.95, p = 0.2137$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LSA</th>
<th>Search</th>
<th>Viz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>$W = 0.80, p &lt; .001$</td>
<td>$W = 0.94, p = 0.08333$</td>
</tr>
<tr>
<td>Topic 2</td>
<td>$W = 0.83, p &lt; .001$</td>
<td>$W = 0.96, p = 0.3994$</td>
</tr>
<tr>
<td>Topic 3</td>
<td>$W = 0.90, p = .015$</td>
<td>$W = 0.91, p = 0.01583$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LDA</th>
<th>Search</th>
<th>Viz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>$W = 0.98, p = 0.891$</td>
<td>$W = 0.94, p = 0.09613$</td>
</tr>
<tr>
<td>Topic 2</td>
<td>$W = 0.93, p = 0.88765$</td>
<td>$W = 0.95, p = 0.1795$</td>
</tr>
<tr>
<td>Topic 3</td>
<td>$W = 0.89, p = 0.009445$</td>
<td>$W = 0.98, p = 0.9116$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dandelion</th>
<th>Search</th>
<th>Viz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>$W = 0.74, p &lt; .001$</td>
<td>$W = 0.54, p &lt; 0.001$</td>
</tr>
<tr>
<td>Topic 2</td>
<td>$W = 0.64, p &lt; .001$</td>
<td>$W = 0.80, p &lt; 0.001$</td>
</tr>
<tr>
<td>Topic 3</td>
<td>$W = 0.85, p = 0.00142$</td>
<td>$W = 0.80, p &lt; 0.001$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wilcoxon rank sum test</th>
<th>H0: Same distribution</th>
<th>Median or Mean (Search, Viz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 1</td>
<td>$W = 460.5, p = .09072$</td>
<td>10632, 9320</td>
</tr>
<tr>
<td>Topic 2</td>
<td>$W = 432, p = .2334$</td>
<td>9595, 7866</td>
</tr>
<tr>
<td>Topic 3</td>
<td>$W = 283, p = .1719$</td>
<td>3564, 3928</td>
</tr>
<tr>
<td>LSA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 1</td>
<td>$W = 508, p = .01189$</td>
<td>.71, .57</td>
</tr>
<tr>
<td>Topic 2</td>
<td>$W = 454, p = .115$</td>
<td>.56, .47</td>
</tr>
<tr>
<td>Topic 3</td>
<td>$W = 298, p = .2693$</td>
<td>.44, .45</td>
</tr>
<tr>
<td>LDA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 1</td>
<td>$t = 0.07, df = 52, p = .9458$</td>
<td>.26 , .26</td>
</tr>
<tr>
<td>Topic 2</td>
<td>$t = -1.04, df = 52, p = .3023$</td>
<td>.18 , .19</td>
</tr>
<tr>
<td>Topic 3</td>
<td>$W = 333, p = .6148$</td>
<td>.201, .200</td>
</tr>
<tr>
<td>Dandelion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 1</td>
<td>$W = 297.5, p = .2632$</td>
<td>.53, .54</td>
</tr>
<tr>
<td>Topic 2</td>
<td>$W = 340, p = .8943$</td>
<td>.58, .58</td>
</tr>
<tr>
<td>Topic 3</td>
<td>$W = 268, p = .1033$</td>
<td>.54, .58</td>
</tr>
</tbody>
</table>

Table 4.2 Task 1 statistics about the four proxies measuring the sensemaking accuracy. The tests’ results as well as the median (non-parametric tests) or mean (parametric tests) are shown. The cells are coloured depending on the tests’ significance.

The Table 4.2 displays all tests’ results (normality tests and comparison tests) related to the first task’s proxies measuring the sensemaking accuracy of the respondents’ answers. All tests, except the semantic LSA for topic 1, significantly accept the null hypothesis.
stating that the two distributions are the same, as outlined in Table 4.2. Therefore, there is no evidence to believe that level of sensemaking generated by the LibSearch users are different than the level of sensemaking generated by the ExploViz users. However, the semantic LSA sensemaking accuracy for topic 1 shows a statistically significant difference between both groups of respondents. In other words, the answers for topic 1 while using the LibSearch interface show a significantly better level of sensemaking accuracy compared to the ExploViz interface. The latter is represented as the cosine similarity metric between the answers and the topic 1 latent semantic spaces.

4.4.3.1.3 Cognitive load measurement

The germane cognitive load corresponds to the aggregation of the experiment’s two first Likert scale questions (“How successful do you think you were [...]” and “How satisfied were you [...]”). The task 1 respondents’ germane cognitive load is significantly different between both interfaces (as outlined in Table 4.1). The ExploViz interface shows a higher germane cognitive load median than the LibSearch, which means the ExploViz interface better supports the first experimental task (Y. Liu et al., 2013). However, the extraneous cognitive load which corresponds to the aggregation of the two last Likert scale questions (“How hard [...]” and “How insecure, discouraged, irritated, stressed and annoyed were you [...]”), accept the null hypothesis stating there is no difference between both distributions. Therefore, the first task does not show a significantly different level of extraneous cognitive load whether the tasks are performed in the LibSearch or the ExploViz interfaces.
4.4.3.2 Task 2 statistics

<table>
<thead>
<tr>
<th>TASK 2 statistics</th>
<th>Shapiro-Wilk normality test</th>
<th>Search</th>
<th>Viz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hover events</td>
<td>$W = 0.82, p &lt; .001$</td>
<td>$W = 0.91, p = .01335$</td>
<td></td>
</tr>
<tr>
<td>Area explored</td>
<td>$W = 0.83, p &lt; .001$</td>
<td>$W = 0.88, p &lt; .001$</td>
<td></td>
</tr>
<tr>
<td>Distance between 2 books</td>
<td>$W = 0.89, p = .01347$</td>
<td>$W = 0.94, p = .1324$</td>
<td></td>
</tr>
<tr>
<td>Exteraneous cognitive load</td>
<td>$W = 0.90, p &lt; .001$</td>
<td>$W = 0.87, p &lt; .001$</td>
<td></td>
</tr>
<tr>
<td>Germane cognitive load</td>
<td>$W = 0.90, p &lt; .001$</td>
<td>$W = 0.87, p &lt; .001$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wilcoxon rank sum test</th>
<th>H0: Same distribution</th>
<th>Median - (Search, Viz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hover events</td>
<td>$W = 281, p = .1598$</td>
<td>16, 36</td>
</tr>
<tr>
<td>Area explored</td>
<td>$W = 1883, p = .8366$</td>
<td>.0017, .0026</td>
</tr>
<tr>
<td>Distance between 2 books</td>
<td>$W = 321.5, p = .4823$</td>
<td>.14, .092</td>
</tr>
<tr>
<td>Exteraneous cognitive load</td>
<td>$W = 1529.5, p = .611$</td>
<td>3, 3</td>
</tr>
<tr>
<td>Germane cognitive load</td>
<td>$W = 1336.5, p = .4706$</td>
<td>3, 3</td>
</tr>
</tbody>
</table>

Table 4.3 Task 2 statistics about the behavioural metrics as well as the germaine and extraneous cognitive loads. The tests’ results as well as the median are shown. The cells are coloured depending on the tests’ significance.

4.4.3.2.1 Level of exploratory behaviour

As outlined in Table 4.3, the three metrics measuring the respondents’ exploratory search behaviour while performing the experiment’s second task are not statistically significant. Therefore, the number of hover events, the area explored and the distance between the two selected books do not show statistically significant difference whether the respondent interact with the LibSearch or the ExploViz interface.

4.4.3.2.2 Cognitive load measurement

As highlighted in Table 4.3, both the extraneous and the germaine cognitive loads show non-significant results from the Wilcoxon test comparing the distributions of the respondents’ Likert scales answers following the second task. Therefore, there is no evidence to state that the ExploViz spatialisation better supports the respondents’ cognition while selecting the two books needed to meet the instructions’ requirements.
4.4.3.3 Task 3 statistics

**Table 4.4 Task 3 statistics about the behavioural metrics as well as the germane and extraneous cognitive loads. The tests’ results as well as the median are shown. The cells are coloured depending on the tests’ significance.**

<table>
<thead>
<tr>
<th>Shapiro-Wilk normality test</th>
<th>Search</th>
<th>Viz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hover events</td>
<td>$W = 0.83, p = 0.001355$</td>
<td>$W = 0.70, p &lt; .001$</td>
</tr>
<tr>
<td>Area explored</td>
<td>$W = 0.76, p &lt; .001$</td>
<td>$W = 0.73, p &lt; .001$</td>
</tr>
<tr>
<td>Exploration time</td>
<td>$W = 0.71, p &lt; .001$</td>
<td>$W = 0.84, p &lt; .001$</td>
</tr>
<tr>
<td>Extraneous cognitive load</td>
<td>$W = 0.87, p &lt; .001$</td>
<td>$W = 0.90, p &lt; .001$</td>
</tr>
<tr>
<td>Germane cognitive load</td>
<td>$W = 0.87, p = 0.001355$</td>
<td>$W = 0.89, p &lt; .001$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wilcoxon rank sum test</th>
<th>H0: Same distribution</th>
<th>Median - (Search, Viz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hover events</td>
<td>$W = 249, p = .1858$</td>
<td>12, 23</td>
</tr>
<tr>
<td>Area explored</td>
<td>$W = 1544, p = .9694$</td>
<td>.0015, .0011</td>
</tr>
<tr>
<td>Exploration time</td>
<td>$W = 334, p = .6271$</td>
<td>141, 161</td>
</tr>
<tr>
<td>Extraneous cognitive load</td>
<td>$W = 1303, p = .3506$</td>
<td>2, 3</td>
</tr>
<tr>
<td>Germane cognitive load</td>
<td>$W = 1431, p = .9057$</td>
<td>4, 4</td>
</tr>
</tbody>
</table>

Table 4.4 Task 3 statistics about the behavioural metrics as well as the germane and extraneous cognitive loads. The tests’ results as well as the median are shown. The cells are coloured depending on the tests’ significance.

4.4.3.3.1 Level of exploratory behaviour

The third task, as presented in Table 4.4, does not show statistically significant difference in the respondents’ exploratory search behaviour. Indeed, the number of hover events, the area explored and the exploration time do not show statistically significant difference whether the respondent interact with the LibSearch or the ExploViz interface, while performing the third task. Therefore, there is no evidence to show that the ExploViz spatialisation supports the user’s exploratory search behaviour when selecting a book from the digital library.

4.4.3.3.2 Cognitive load measurement

As highlighted in Table 4.4, both the extraneous and the germane cognitive loads show non-significant results from the Wilcoxon test comparing the distributions of the respondents’ Likert scales answers following the third task. Therefore, there is no evidence to state that the ExploViz spatialisation supports the respondents’ cognition while selecting a book from the digital library.

4.5 Summary and discussion

The between-group online task-based experiment is designed in order to capture information about a potential difference in respondents’ behaviours while performing
three different types of exploratory search tasks (sensemaking exploration, directive exploration and free exploration). The experiment collected 54 clean and complete answers which were analysed using a set of novel metrics (discussed in section 4.1.5). The latter are designed to express levels of exploratory search behaviours and levels of sensemaking while performing the experimental tasks. Additionally, to the behavioural and sensemaking metrics, a set of 4 questions is systematically asked after each task. Two questions measure the germane cognitive load indicating how cognitively supportive the underlying tool is and the two other questions measure the extraneous cognitive load indicating levels of cognitive overhead generated by the underlying tool.

The Table 4.5 summarises all p-values resulting from the tests comparing distributions resulting from the LibSearch and the ExploViz interfaces while performing the three experimental tasks. The green fields outline the significant differences supporting the project’s hypotheses, with a confidence level of 5%. The red field highlights the significant difference opposing the project’s hypotheses, with a confidence level of 5%. All the other fields showing ns express a non-significant difference in the respondents’ behaviours, cognitive loads or sensemaking levels when performing the tasks in whether the LibSearch or the ExploViz interfaces. One should notice that the statistical differences found are slightly significant ($p < .05$) and that they only appear in the first task measuring the respondents’ sensemaking.
Table 4.5 Summary of all tested p-values comparing distributions resulting from the LibSearch and the ExploViz interfaces while performing the three experimental tasks

<table>
<thead>
<tr>
<th>Respondents' behaviour</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hover events</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Area explored</td>
<td></td>
<td>* (57e-4, 81e-4)</td>
<td>ns</td>
</tr>
<tr>
<td>Distance between 2 books</td>
<td>-</td>
<td>ns</td>
<td>-</td>
</tr>
<tr>
<td>Exploration time</td>
<td>-</td>
<td>-</td>
<td>ns</td>
</tr>
<tr>
<td>Cognitive Load</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germane</td>
<td></td>
<td>* (3, 4)</td>
<td>ns</td>
</tr>
<tr>
<td>Extraneous</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sensemaking accuracy</th>
<th>Task 1</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic proxy</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Semantic LSA proxy</td>
<td></td>
<td>* (.71, .57)</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Semantic LDA prox</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Semantic Dandelion proxy</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- Not computed
ns (p ≥ .05 (non-significant))
* (Mdn LibSearch, ExploViz) (p < .05)
** (Mdn LibSearch, ExploViz) p < .01

Given the low levels of significance and the high number of largely non-significant tests, the possible conclusions derived from task 1 should be carefully interpreted. The first task’s results show that the respondents explored more of the information space but with the same number of hover events, which can be interpreted that these respondents conducted a more efficient exploration process. Indeed, given the same action, the ExploViz’ users looked at a bigger variety of books by covering a bigger information surface. Moreover, the results also show a significantly higher level of germane cognitive load which express the respondents’ perception of their success and satisfaction while performing the first task. The higher the germane cognitive load the more supportive an exploratory search tool (Y. Liu et al., 2013). However, one of the four proxies designed for measuring the users’ sensemaking accuracy shows a significantly lower level of accuracy when describing topic 1 than the LibSearch equivalent interface. This result represents one twelfth of all other similar tests measuring sensemaking accuracy, which all lead to non-significant results. Therefore, no strong conclusion can be provided towards a potential difference between both interfaces. One can thus conclude that both interfaces provide similar support towards sensemaking accuracy.
5 CONCLUSION

5.1 Introduction

The research project has focused on the comparison between a visual search interface, namely ExploViz and its non-visual equivalent, namely LibSearch at their ability to support exploratory search within a digital library. The project addressed an aspect of visualisation research that is not well supported in the literature. More specifically, the research question focused on the potential impact of visual spatialisation of an underlying collection of documents on the users’ exploratory search behaviour while performing exploratory search tasks. This required the implementation of two distinct search interfaces - one visual and one non-visual, which could be quantitatively compared in an exploratory task-based experiment. The visual search interface was implemented and justified within a controlled visualisation framework and derived from design considerations drawn from a review of the literature. The non-visual interface was implemented with respect to the state-of-the-art keyword-based search interfaces.

Following the visualisation design study and both interfaces’ implementations, a between-group online experiment was configured in order to explore the comparison of both search interfaces. They both are data equivalent and usable in an exploratory search context. Three user-tasks, meeting the criteria of exploratory search, were designed. The data collected by the online experiment include several proxies for measuring sensemaking accuracy, several metrics capturing the exploratory aspect of the respondents’ behaviour and two facets of the respondents’ cognitive loads (germane and extraneous).

The user-tasks and the metrics were thoroughly designed so that the respondents performing the experiment enter a process of exploratory search of which exploratory aspects could be measured. The results’ methodology consisted in comparing both groups’ results, each corresponding to one search interface. The sensemaking task’s results show that using the ExploViz interface significantly increased the respondents’ area explored and germane cognitive load, and significantly decreased one indicator of sensemaking accuracy for topic 1, compared to the LibSearch interface. However, the two other exploratory search tasks’ results show that using the ExploViz interface does not influence exploratory search behaviour nor cognitive load when compared with the LibSearch interface.
The project shows mixed results that must be interpreted with care. On the one hand, the respondents’ exploratory behaviour showed no clear difference whether interacting with a spatial representation of a digital library's content or with a list-based search view. On the other hand, an improvement in area explored and germane cognitive load while performing a sensemaking task in the ExploViz interface is an encouraging result for further research about visual support for exploratory search tasks within large collections of documents.

5.2 Limitations
The majority of the tasks performed by both independent groups do not exhibit statistically significant behaviour. Those results cannot be directly related and compared to other results from the literature as the ExploViz experimental design shows a greater degree of novelty. However, the literature has generally observed supportive behaviours from information visualisation towards exploratory search. Several limitations might have impacted the experimental design and might explain why the presented hypotheses cannot be statistically verified.

Firstly, the exploratory search tasks are challenging to design. Even though they have been designed with respect to the well-defined exploratory search, they did not incorporate the time constraint as defined by Ryen W. White & Resa A. Roth (2009): “Exploratory search sessions can transcend multiple query iterations and potentially multiple search sessions. An exploratory search can last for days, weeks, or months depending on the nature of the search task [...]” (p. 21). The fast implementation and online context of the experimental design was a compelling factor in regards to the tasks’ design. Indeed, asking for long lasting, multiple sessions and complex tasks would probably have generated a higher dropout rate.

Secondly, the inability to check whether respondents entered an exploratory search behaviour or not could be considered as a limitation. Indeed, there is no certainty about the precise type of behaviour demonstrated by all respondents. For example, the complexity of the underlying technical documents might justify a loss of interest and curiosity leading the respondents to perform the tasks too quickly. Although the complexity of the underlying information space should enhance exploratory search behaviour, the respondents’ interest might as well be too low to explore. The right balance between curiosity and complexity is difficult to evaluate and measure. It is
difficult to know the extent to which the respondents are trained and accustomed to performing exploratory search tasks using lookup search strategies. Even though Athukorala, Glowacka, Jacucci, Oulasvirta, & Vreeken (2015) defined indicators, they can only be applied on keyword-based query-search interface type.

Finally, users are accustomed to the traditional lookup search interfaces for which they probably developed tactics to meet their exploratory search needs. Consequently, there might possibly exist a bias in favour of the traditional search interface resulting from the contrast between a widely adopted interface and a novel interface.

5.3 Review of thesis contribution
Chapter 2 contributed to providing a set of design considerations which can be used by visual designers to implement a visual search interface focused on the users’ exploratory search.

Chapter 4 contributed to providing a set of quantitative metrics for evaluating sensemaking accuracy and exploratory search behaviour, which captured information about a significant difference in task 1 respondents’ exploratory search behaviour between the two search interfaces.

Chapter 4 also contributed to providing an empirical study to assess the impact of a digital library’s visual spatialisation on users’ exploratory search behaviour compared to traditional list-based search interface. The experiment is generalizable so that it can be reproduced independently and used as a benchmark for further investigation¹⁶.

5.4 Reflections and Future research
With the benefit of hindsight and following several feedbacks highlighting the complexity of the library’s content, it might have been more efficient to leverage the *diversive* curiosity factor which leads to exploratory search behaviour (Pace, 2004, p. 343), rather than the complexity factor which can potentially turn into a loss of interest preventing exploratory search behaviour. Given the time required to perform the experiment and the small amount of metadata available in both interfaces, an easy but inspiring subject such as films, cooking, travelling, sports might have been a good way to enhance exploratory search behaviour.

¹⁶ All codes are available on [http://ditviz.com](http://ditviz.com).
Keyword-based search interfaces have remained the overall predominant operating interface for quick retrieval of information when the users can formulate their search goals. Additionally, the recommender systems have been experiencing a significant development and could be described as a tool meeting the users’ unknown or unidentified needs by providing casual browsing. However, the suggested exploratory search experience remains confined to a list-based interface preventing from overall sensemaking and serendipity. The project approached two types of representations: the traditional list-based search results (LibSearch) and the entire library space representation (ExploViz). One avenue for further investigation could contrast these two spaces with a recommendation-based visualisation where recommended paths and recommended aggregation and filters could shape and support the users’ exploration. This would be coherent in today’s context where recommender systems and hybrid interfaces are more widely used and yet often limited to a list-view presentation where items’ locations on the page do not hold information about their relationships among others.
6 BIBLIOGRAPHY


McKay, D., & Conyers, B. (2010). Where the streets have no name: how library users get lost in the stacks. In *Proceedings of the 11th International Conference of the*


7 APPENDICES

7.1 Results’ distributions boxplots

7.1.1 Proxies for sensemaking accuracies across topics and interfaces

The following boxplots present the four distinct types of sensemaking accuracies across topics and interfaces.

7.1.1.1 Syntactic proxy for sensemaking accuracy

![Boxplot of syntactic sensemaking across topics and interfaces](image)

*Figure 7.1 Syntactic sensemaking across both interfaces (search corresponds to the LibSearch distribution and Viz corresponds to the ExploViz distribution) and per topic*
7.1.1.2 Semantic LSA proxy for sensemaking accuracy

![Semantic LSA Sensemaking across Topics and Interfaces](image)

*Figure 7.2 Semantic LSA sensemaking across both interfaces (search corresponds to the LibSearch distribution and Viz corresponds to the ExploViz distribution) and per topic*

7.1.1.3 Semantic LDA proxy for sensemaking accuracy

![Semantic LDA Sensemaking across Topics and Interfaces](image)

*Figure 7.3 Semantic LDA sensemaking across both interfaces (search corresponds to the LibSearch distribution and Viz corresponds to the ExploViz distribution) and per topic*
7.1.1.4 Semantic Dandelion proxy for sensemaking

Figure 7.4 Semantic Dandelion sensemaking across both interfaces (search corresponds to the LibSearch distribution and Viz corresponds to the ExploViz distribution) and per topic

7.1.2 Respondents’ exploratory behaviour across tasks and interfaces

The following boxplots present the four distinct indicators measuring levels of the respondents’ exploratory behaviour across tasks and interfaces.

7.1.2.1 Number of hover events

Figure 7.5 Number of hover events across tasks and interfaces (search corresponds to the LibSearch distribution and Viz corresponds to the ExploViz distribution)
7.1.2.2  Area explored

![Box plot of Area explored across Tasks and Interfaces](image1)

Figure 7.6 Area explored measured as the surface covered by each topic’s polygon. The results are split by tasks and by type of interface (search corresponds to the LibSearch distribution and Viz corresponds to the ExploViz distribution).

7.1.2.3  Distance between two selected books

![Box plot of Distance between two selected books across Interfaces](image2)

Figure 7.7 Euclidean distance between two selected books from the experiment's task 2, across interface types (search corresponds to the LibSearch distribution and Viz corresponds to the ExploViz distribution).
7.1.2.4 Exploration time

![Exploration time (in seconds) across Interfaces](image)

Figure 7.8 Exploration time of task 3 extracted in seconds while the experiment’s task 3 is performed, per interface type (search corresponds to the LibSearch distribution and Viz corresponds to the ExploViz distribution)

7.1.3 Cognitive load measurement across

The bar charts present the distributions of the 5 points Likert scale answers, following each experimental tasks.

7.1.3.1 The respondents’ germane cognitive load

Figure 7.9 shows the aggregation of the two first Likert scale questions (“How successful do you think you were [...]” and “How satisfied were you [...]”). The faceted bar chart shows the percentage of the five point Likert scale answers per task (the 3 tasks are shown on the right) and per interface type (the x axis). The number of respondents between the ExploViz and the LibSearch being different, the percentages are computed based on the number of answers found in each task and for each type of interface. The closer to 5, the higher germane cognitive load and the better cognitive support.
7.1.3.2 The respondents’ extraneous cognitive load

Figure 7.10 shows the aggregation of the two last Likert scale questions ("How hard [...]" and "How insecure, discouraged, irritated, stressed and annoyed were you [...]”). The faceted bar chart shows the percentage of the five point Likert scale answers per task (the 3 tasks are shown on the right) and per interface type (the x axis). The number of respondents between the ExploViz and the LibSearch being different, the percentages are computed based on the number of answers found in each task and for each type of interface. The closer to 5, the higher extraneous cognitive load and the worse cognitive support the interface is in the response process.
7.2 Resulting visual spaces of Document-Term Matrix initial transformations

Five distinct weighting schemes are tried:

- Term frequency weighting (TF): w0
- Binary(DTM) * Inverse document frequency (IDF): w1
- Log(TF) * term’s entropy: w2
- TF * IDF: w3
- Log(TF) * IDF: w4

The Euclidean and the cosine distance are computed on each weighted term document matrix. Additionally, the Jaccard index is applied on the binary matrix only. The resulting matrices are square similarity Mii where i = 8242 documents. The cosine similarity matrices are manipulated in order to reflect a distance (the diagonal is removed and Mii <- 1 – Mii, so that the distance ranges from 0 to 2.

Then multidimensional scaling (MDS) is applied on each similarity matrix. Finally, different clustering techniques are performed in order to apply a colour on each document.

- k-means clustering is performed on the sparse dtm with 5 clusters
• k-means is also tried on the binary sparse adjacency matrix reflecting the number of terms each pair of documents have in common
• walktrap community algorithm is applied on the adjacency matrix

The above transformations therefore result in a table which assigns a classification and coordinates to each book. That information is encoded into a spatialisation and a colour code classification. A selection of the resulting visual representations is presented below.

Figure 7.11 Preview spatialisation from R, ggplot2. Early transformation: w3_dtm + cosine_distance + 4Walktrap_community. The cosine distance is performed on the TF * IDF document term matrix. The colours represent the 4 walktrap communities computed independently on the adjacency matrix.
Figure 7.12 Preview spatialisation from R, ggplot2. Early transformation: w1_dtm + euclidean_distance + 5 Clusters Kmeans: The Euclidean distance is performed on the Binary(DTM) * Inverse document frequency (IDF). The colours represent the 5 k-means clusters based on the binary adjacency matrix.
Figure 7.13 Preview spatialisation from R, ggplot2. Early transformation: w0_dim + jaccard_index + 5 Clusters_Kmeans: The Jaccard index is performed on the Term frequency weighting matrix. The colours represent the 5 k-means clusters based on the binary adjacency matrix.
Figure 7.14 Preview spatialisation from R, ggplot2. Early transformation: w2_dtm + cosine_distance + 5Clusters_Kmeans: The cosine distance is performed on the log(TF) * term’s entropy DTM. The colours represent the 5 k-means clusters based on the binary adjacency matrix.