# 1 Introduction

The last decade has experienced an exponential growth of the amount of publicly available data, mostly leveraged by the spreading of the WWW, the development of semantic technologies, and the Linked Data community[1]. Since the mass publication of semi-structured information has taken place, the search for information has been one of the most relevant themes in information systems research. Unfortunately the development of efficient applications to empower the searcher in finding and processing relevant pieces of information has not evolved with the same speed, leaving large parts of the published data effectively out of reach.

It has been a consensus among the research community that traditional Information Retrieval (IR) tools are not sufficient in supporting data exploration behavior (BATES, 1989; KUHLTHAU, 1991; MARCHIONINI, 2006; WHITE *et al.*, 2007), which takes place when the user lacks the knowledge and ability to precisely describe the characteristics of the desired items in terms of keywords. The interaction in IR tools, such as Google, and Yahoo!, is based on isolated sequences of query-response actions, where the user formulates a query and the system returns a set of documents that match the query. Despite the simplicity of use of the query-response model, it has been strongly criticized with regards to supporting more complex information search tasks, such as Exploratory Searches (MARCHIONINI, 2006). As an example, the task "discover who invented the light bulb" can be solved in a single query against search engines. On the other hand, the task "write a paper about recently discovered treatments for diabetes" would require exploratory behavior and, consequently, more advanced support for browsing, filtering, aggregating, and comparing information items. In fact, traditional IR designers have already recognized the need to support more complex information searches and question-answering tasks in their tools. For example, Google and Microsoft search engines already provide result set

[1] http://linkeddata.org/

[2] Code available at https://github.com/trnunes/xplain and demonstrations

browsing and filtering operations based on semi-structured data, leveraged by their knowledge graphs (SINGHAL, 2012). The support for exploration tasks, though, is still in its infancy.

Information exploration environments aim at supporting information gathering tasks that often involve a high degree of complexity, lack of user's knowledge about the data, multiple items and data types, and that do not have a clear ending (WILDEMUTH; FREUND, 2012). Such tasks usually involve sequences of data interactions beyond the query-response paradigm that eventually lead to the desired outcome, which can be either a set of items or a knowledge state acquired along the process. Given the complexity of the tasks, exploration environments should be carefully designed to support a rich enough variety of data processing actions and strategies, accessed through their interfaces.

Even though there has been much work on the development of computational systems supporting exploration tasks, such as visualization systems, faceted search tools, and set-oriented interfaces, the lack of a formal understanding of the exploration process and its operations and the absence of a proper separation of concerns approach in the design phase is the cause of many expressivity issues and serious limitations. The exploration environments usually do not present precise descriptions of neither what operations they support nor what combinations of operators are available through their interfaces. Thus, it is not possible to assess how good is the support of an environment for a given exploration task from the functional perspective. As a consequence, no matter how advanced the interface design is, a missing operator may cause serious limitations for the task resolution process and the range of distinct solution strategies the explorer can adopt.

Moreover, exploration strategies can also be a valuable source of information, where future explorers can draw upon the experience acquired in previous explorations. The support for reuse of exploration strategies is also absent in the majority of the state-of-the-art tools.

This work presents a novel formalization of exploration processes and operations that is able to describe at least the majority of exploration environments of the state-of-the-art. The proposed framework of exploration operations leveraged a new design approach for environments where functionality and interface issues are addressed separately. Moreover, we present a new exploration system that generalizes the majority of the state-of-the art exploration tools. The evaluation of the proposed framework is guided by case studies and comparisons with state-of-the-art tools. The results show the relevance of our approach both for the design of new exploration tools with higher expressiveness, formal assessments and comparisons between different tools, and to promote reuse of exploration solutions in new explorations.

## 1.1.Research Questions

Although the exploration phenomenon has been studied along the last decade under the concept of "Exploratory Search" (WHITE *et al.*, 2007), no conclusions have been drawn with regards to a sufficient set of actions involved in the process. Therefore, the central question addressed by this research is: can we characterize Information Exploration behavior in a precise way? A second derived question is, is there a sufficient enough set of primitive exploration actions that can support most exploration tasks?

Considering that exploration actions can be modeled as data manipulation operations, more specific questions arise. First, which operations are involved in an exploration process? What are their parameters? What are the results of their application to a set of items? Second, an exploration is accomplished by a sequence of actions where the results of previous actions can be used as input to subsequent actions, hence, forming functional compositions of data manipulation operations. Thus, which compositions can be formed? For example, is it possible to issue a query over the results of a refinement operation?

By providing answers to these questions we expect to build a unified framework of exploration operations, which we believe can benefit the whole area of information exploration in the following ways:

- Aggregating the knowledge and findings concerning data manipulation operations in exploration tasks in a common framework of operations;
- Leveraging the design process of expressive exploration tools by promoting a separation of functional concerns from interaction and interface concerns;
- It can be used as a framework for analyzes and comparisons of exploration tools, where designers and researchers can qualitatively

assess the extent of the tools support for exploration tasks under the same set of operations;

- Since some exploration tasks are recurrent within communities of information consumers (MUKHERJEA; BAMBA; KANKAR, 2005; SHIH; LIU; HSU, 2010), we expect that the formal descriptions of explorations would leverage the discovery and reuse of exploration patterns in future explorations.

## 1.2. Research Goals and Methodology

The main goal of this research work is to find and formalize a sufficient set of actions that allows the description and representation of exploration processes and more precise comparisons of state-of-the-art tools. However, we do share the vision that achieving a definitive framework, if possible, is a very hard task (WILSON, MAX L.; SCHRAEFEL; WHITE, 2009). Accordingly, our formal model aims at describing a representative set of state-of-the-art operations, rather than all possible actions, as this is an open-ended question.

In order to ensure a satisfactory coverage of the framework, we first carried out deep analyzes of state-of-the-art publications describing tools, tasks, and behavioral models. From the tools we extracted a collated list of features. From this list, we abstracted out interface and interaction details to derive abstract functions, with their parameters and return values. We evaluated the expressivity of the framework using case studies and comparisons with state-of-the-art tools.

The second goal is to leverage the design of new and expressive exploration environments, where the functionality is addressed separately from interaction and interface concerns. In order to achieve this goal, we devised a new approach to the design space of exploration environments based on the formalization of the functional aspects, which led to a three-layered architecture. Moreover, we discuss the issues and solution alternatives within each layer in terms of the design and development of a new exploration environment XPlain[2].

[2] Code available at https://github.com/trnunes/xplain and demonstrations available at https://www.youtube.com/watch?v=ipgue99RZcw and https://www.youtube.com/watch?v=VkzE2ONZNWA&t=7s

As part of the methodology, both the framework and its application to real exploration cases, evaluations, and also to the design of exploration environments were discussed and validated among the research community in specialized workshops, such as the works in (NUNES; SCHWABE, 2014, 2015, 2016) published on the International Workshop on Intelligent Exploration of Semantic Data (IESD), and the work in (NUNES; SCHWABE, 2017) published on the Visualization and Interaction for Ontologies and Linked Data (VOILA).

This thesis is organized as follows. Chapter 2 presents a discussion about the characteristics of exploration tasks and processes. Chapter 3 introduces a framework of reference to approach the design space and issues of exploration environments. We present a formalization approach of exploration processes and actions in chapter 4. Chapter 5 presents case study evaluations of the proposed framework of operations. In chapter 6 we present a new evaluation approach of the expressivity of exploration environments, leveraged by the formal framework of operations. Chapter 7 presents a discussion of the design space of exploration environments in terms of the development of the new environment XPlain. Chapter 8 presents the research conclusions and discusses future works.