Research project - proposal

Study programme: Software and data engineering

Project type: research project

Student: Quoc Anh Nguyen

Supervisor: Martin Nečaský

Consultants: Jakub Klímek, Petr Škoda, Štěpán Stenchlák

Project name: Helping users navigate data specification

Introduction:

Dataspecer is a tool designed for managing and modeling data specifications. It helps users create structured data schemas from ontologies and export them into various formats such as JSON, XML, and CSV.

This project aims to enhance the Dataspecer tool with a standalone application, which will allow users to ask queries using natural language and receive their query in a technical language - SPARQL. Moreover, this proposed extension will help guide users through the data specification defined in Dataspecer.

Motivation:

In an organization, there are usually only a handful of people who fully grasp the whole domain ontology. My proposed application will help users, who do not have technical background nor full understanding of the domain ontology, navigate the ontology and ask about what they need. It is often the case that users do not realize what they could be asking about until they are presented with the possibility. The proposed application will hint the users about useful properties contained in the data specification so that they can decide whether or not it is of interest to them.

Project description:

I will join the research team that works on the Dataspecer tool and implement an application whose main purpose is to help users navigate the data specification they have selected. The user will choose a data specification that they want to query about and input their initial query in natural language. The application will translate the user's query into a SPARQL query and at the same time will offer properties related to the user's query which might interest them. A short summary will be displayed for each offered property. The user can choose to use some or all of the offered properties to refine their query. The application then responds by adding the selected properties to the user's current query and offering more properties, if possible. The interaction between the user and the application will continue in an iterative manner for as long as the user desires.

The main focus of this project will be to explore the possibilities that large language models have to offer in order to map concepts, which are in users' queries, yet obscured by natural

language, to concepts in the data specification. At the same time I will use LLMs to find suitable properties in the data specification to offer to users for possible query expansion and to summarize concepts from the data specification in a way such that users with non-technical background will be able to understand them.

It is important to note, that the development of large language models is moving ever forward and therefore I will design the proposed application to be independent of any specific LLMs and any specific prompting strategies. Prompt templates will be sufficiently separated from the code to ensure easy configuration. I will choose a few language models for the purposes of development and testing, but the application itself will not be tied to them.

Project output:

The output of the project will be a standalone application, which will function independently from the Dataspecer tool. This application will take a data specification or an URL to a data specification and then allow users to query in natural language.

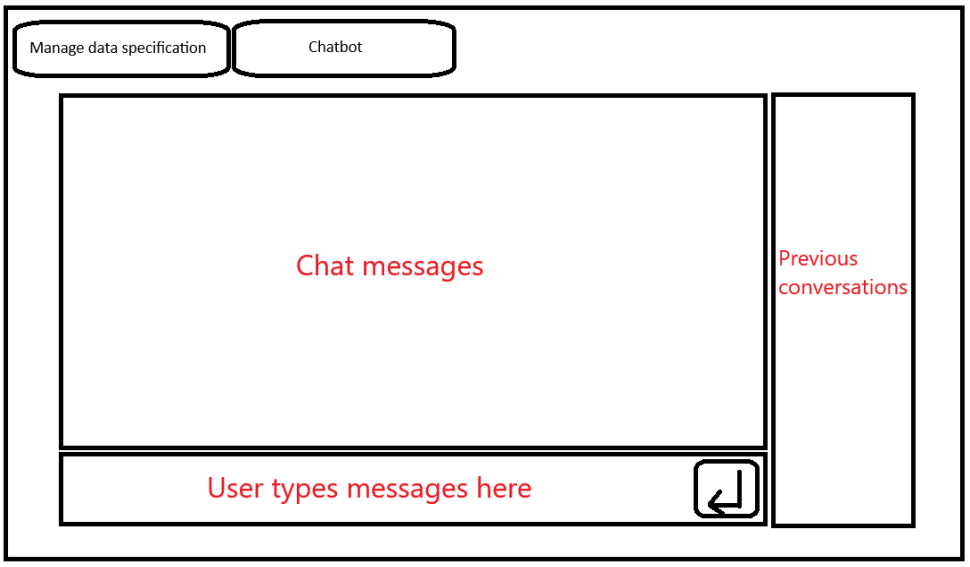
The application will compose of two parts: the front end and the back end.

Users will interact with the front end of the application, which will be a web page. The front end will send user input to the back end for further processing.

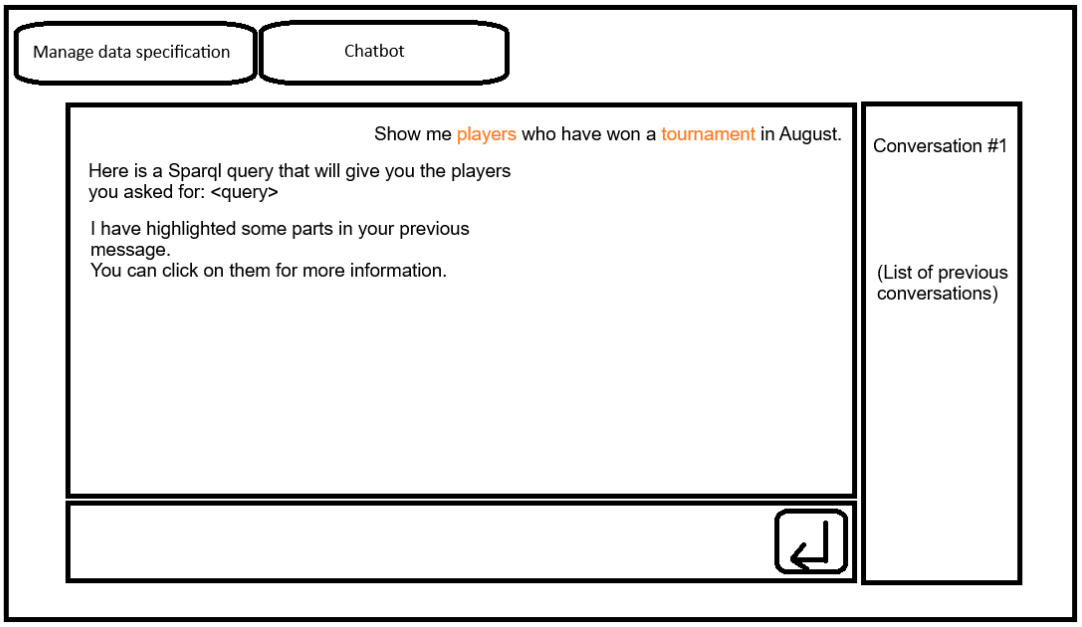
The back end will have a modular architecture in order to achieve separation of concerns. One module will handle communication with the front end, another module will be responsible for communicating with the Dataspecer API, and another for sending prompts to the LLM and retrieving and processing answers from the LLM. The back end will also have a database where the application will store a history of users' conversations.

This project will integrate into the Dataspecer tool by extending its data specification manager with an additional functionality. Each data specification will have a button that uploads that specification into my standalone application and redirects users to the UI of my application. Users will then be able to query over the data specification that they have selected.

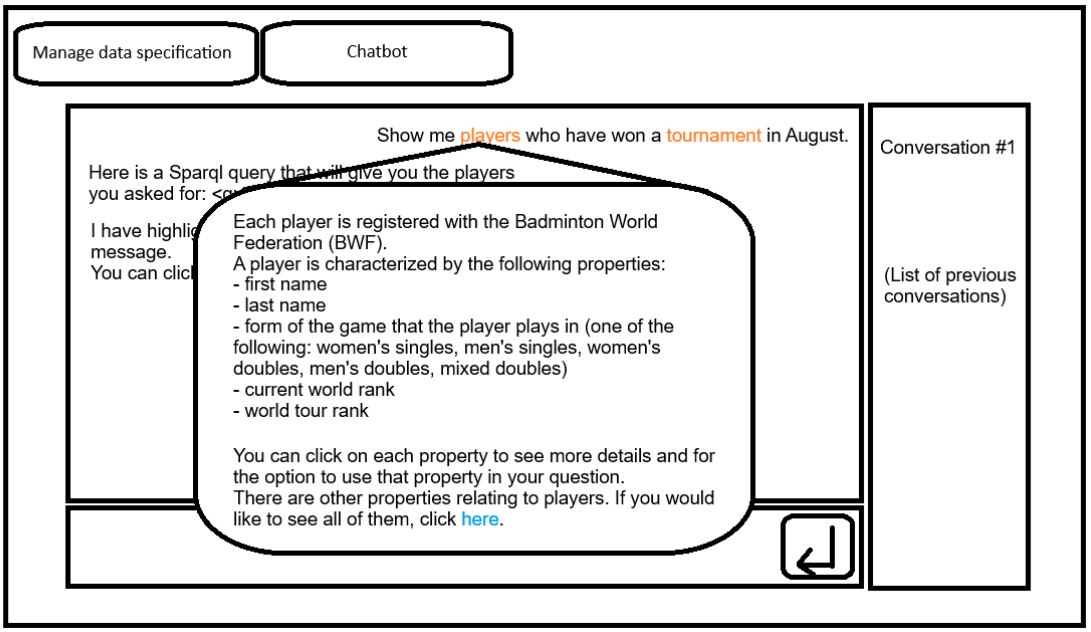
Preliminary front end mockups



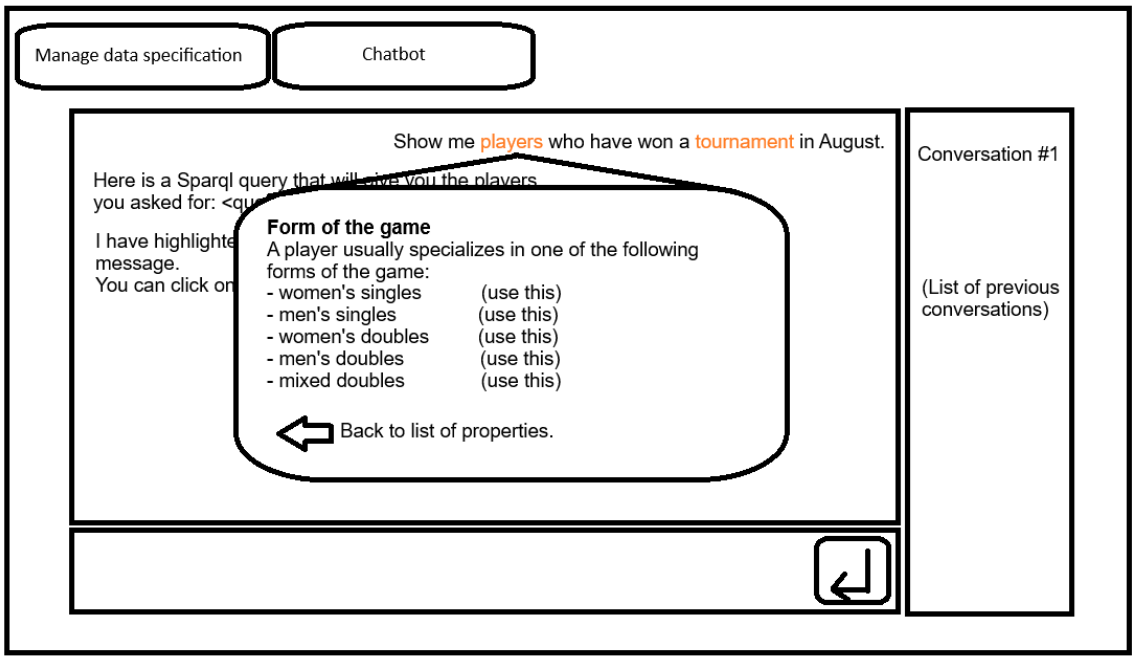
The UI could look something like the picture above. It is a classic chatting interface with one added element - the “previous conversations” part on the right. Users type messages into the bottom box. They can see the messages in the box above that. The area labeled “previous conversations” on the right side contains the current conversation and other conversations that users have had. Users will be able to select a conversation and continue in it.



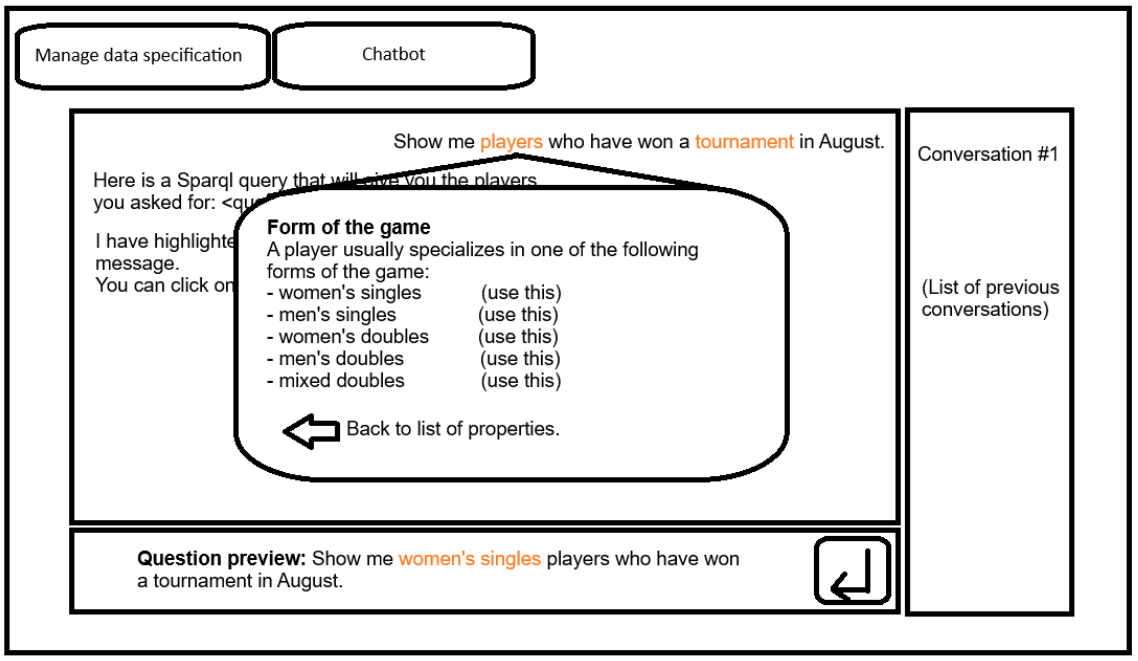
The user’s messages are displayed on the right side and the system messages on the left. Words in the user’s question, which have been mapped to some concepts in the data specification, will be highlighted. In this case, those are the words “players” and “tournament”.



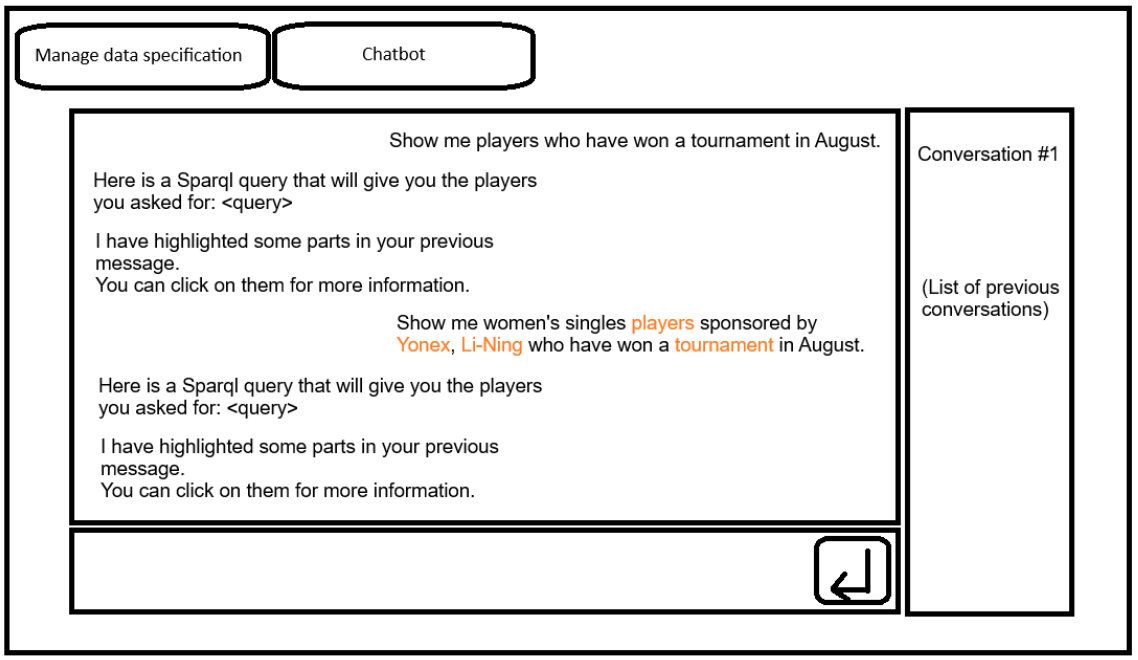
Clicking on a highlighted word will display a short summary for the concept it represents.



Users can click on a property in the concept's summary and the system will display information about that property. Users can choose to use that property to expand their original question.



The system will display a preview of what the expanded question would look like if the property were to be used.



The same process happens for the expanded questions. This goes on until the user stops the conversation.