Clean Coders

UNIVERSITY OF TEXAS AT DALLAS | 800 W CaMPBELL RD RICHARDSON TX 75080

FINAL REPORT

Semantic web



Table of Contents

[Group Information 2](#_Toc435999118)

[Details of team members 2](#_Toc435999119)

[Report 3](#_Toc435999120)

[1. Introduction 3](#_Toc435999121)

[2. Target Audience 3](#_Toc435999122)

[3. Description of data sources 3](#_Toc435999123)

[4. Data Integration 4](#_Toc435999124)

[5. Data Product Results 5](#_Toc435999125)

[6. Custom project Justification 9](#_Toc435999126)

[Summary 10](#_Toc435999127)

# Group Information

Group Name: Clean Coders

Project Name: Relation extraction using semantic web technologies

Type: Custom Project

Team Members: 4

## Details of team members

1. Ekal Golas – exg140230
2. Navaneeth Rao – nbv140130 (In a different section - 502)
3. Tejaswini Sirlapu – txs142130
4. Prashanth Govindaraj – pxg142030

# Report

## Introduction

This project is about relation extraction from unstructured business data using semantic web technologies. It aims to derive unstructured data from the web and process content in the data to find entities and relationships. These entities and relationships, modelled by a semantic ontology, organizes the content in the data to a structured format. The results are then displayed on to a website, where the user can search for different entities and obtain relationships and linked information about that entity.

This report discusses the target audience for this project. Then it explains the data and discusses about the custom ontologies used to describe the structured data. After that, some focus is laid on the data integration and its value and then some results are displayed.

The report finally justifies the custom project choice and concludes with a brief summary.

## Target Audience

This project can be relevant/useful to the following:-

1. Financial/Marketing analysts – As a relation between entities and the money is developed, the structured data can be used to answer queries based to money involved with a person or an organization
2. Statistical users – This structured data can be queried to obtain statistics around a region, person, organization or location
3. Random web user – A random web user can browse this website instead of browsing the web to find unstructured information, and find relationships and linked information about a person or an organization in one place.

## Description of data sources

The data is described as:-

1. A collection of web pages, indexed and stored in SolR
2. Data will be about people, organizations, locations and related financial information
3. Initially we took a small sample of test data, and then scaled out. For this report, the sources used are:-
   1. <https://en.wikipedia.org/wiki/Bill_Gates>
   2. <http://www.biography.com/people/bill-gates-9307520>
   3. <http://www.bloomberg.com/news/articles/2015-11-22/bill-gates-hopeful-china-will-amend-laws-to-boost-philanthropy>
4. Data is available publicly on the internet

The following ontology describes the structure in which data will be described in the results:-

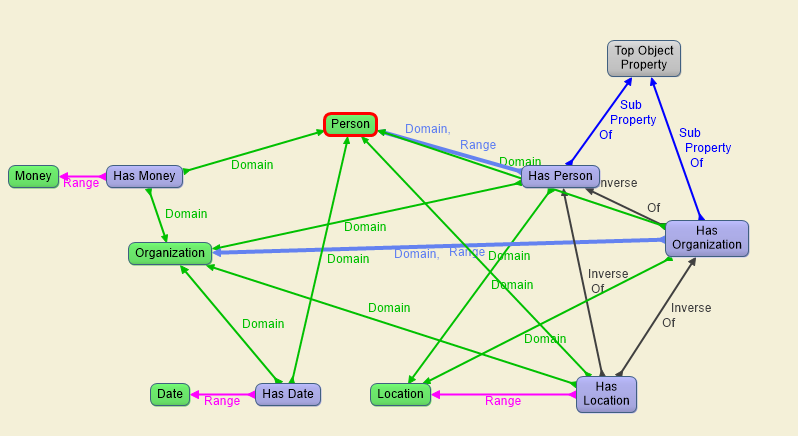


Figure 1: Custom Ontology Description

## Data Integration

Data is integrated in the following steps:-

1. Firstly, a set of documents and content is stored in Apache SolR
2. Then this data is read from SolR and fed into Stanford CoreNLP API to tag the data with various properties
3. The entity tags and related data are then fed into Apache Jena, which reads the custom ontology in Figure 1 above as RDF/XML, and transforms the data tagged with entities into structured data stored a RDF and N-triples files
4. These files are then read into a website and the results are displayed to the user

This integration from end-to-end provides the experience of transformation of unstructured data scattered over the web into a structured, linked format that the user can make sense of and derive meaningful information from.

## Data Product Results

Following was the expected results of this project:-

1. Getting an insight as to how people, organization and locations are related to each other.
2. Possible financial information related to these results
3. Visualization of this data as a graph

These results are interesting because scope of this project can help marketing teams make better sales/business decision based on extracted relationships and structures data visualizations. Even casual users might benefit by gather general knowledge from this data about companies and famous personalities.

A sample visualization graph looks as follows:-

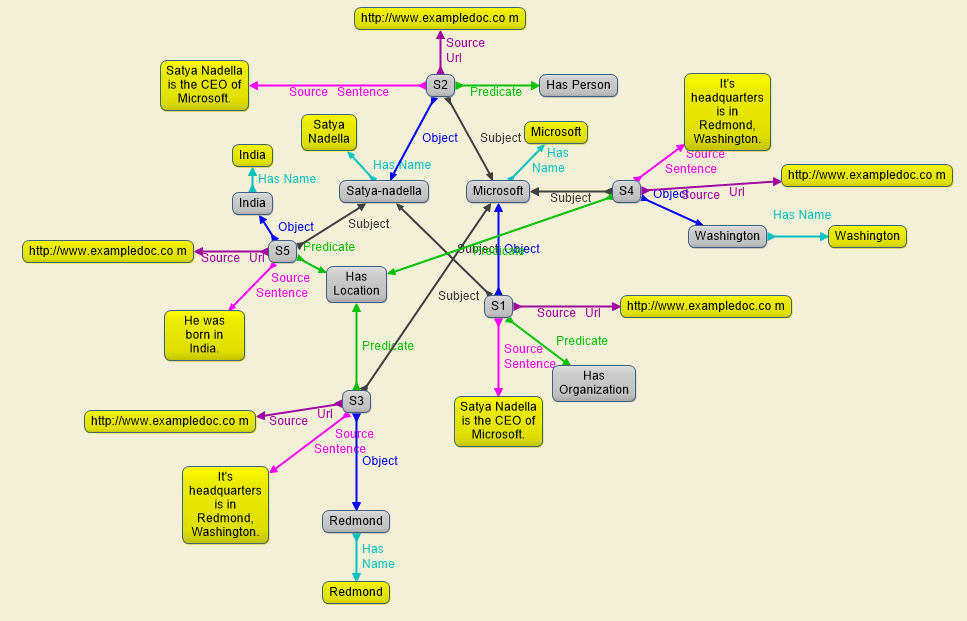


Figure 2: A Sample Visualization Graph

The above graph stored as N-triple format looks as follows:-

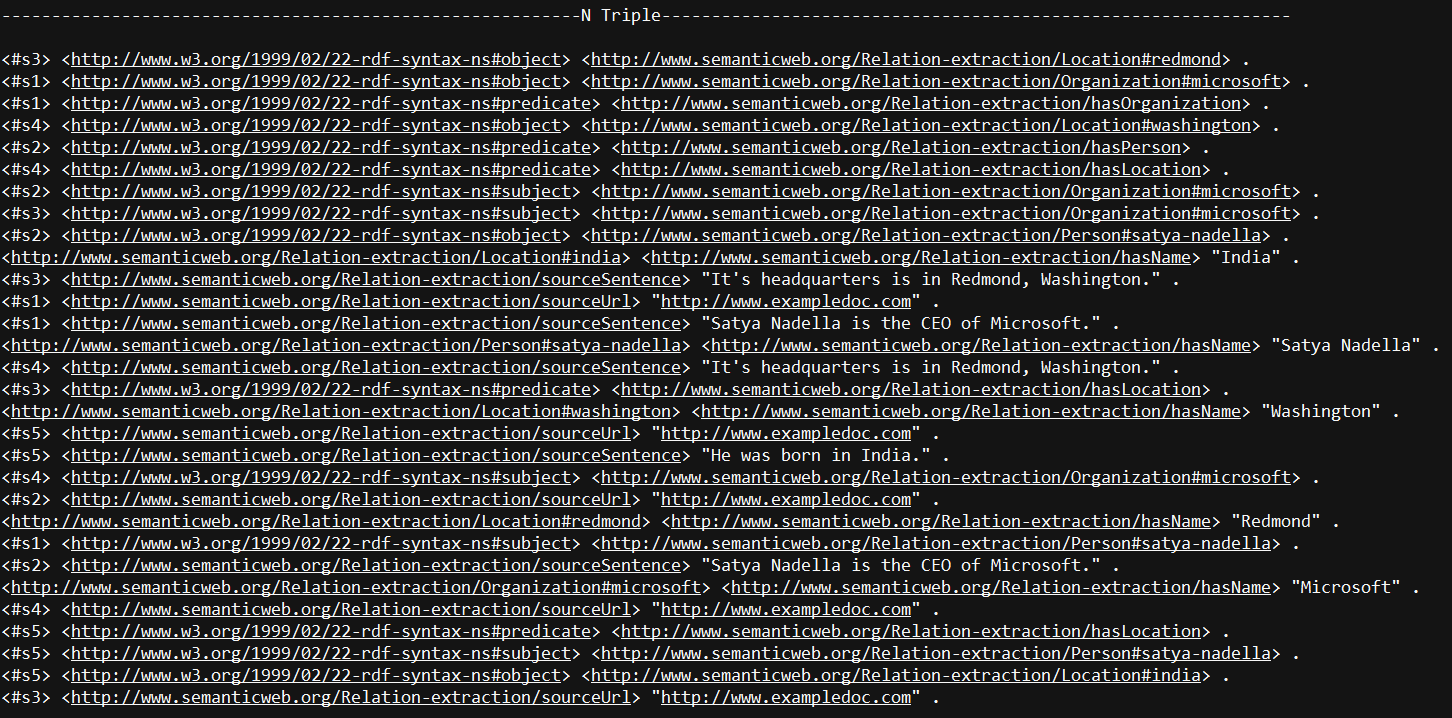


Figure 3: Data In N-triple Format

The website will display results about an entity on the linked dates, organizations, money and locations. A sample results displayed related dates is as follows:-

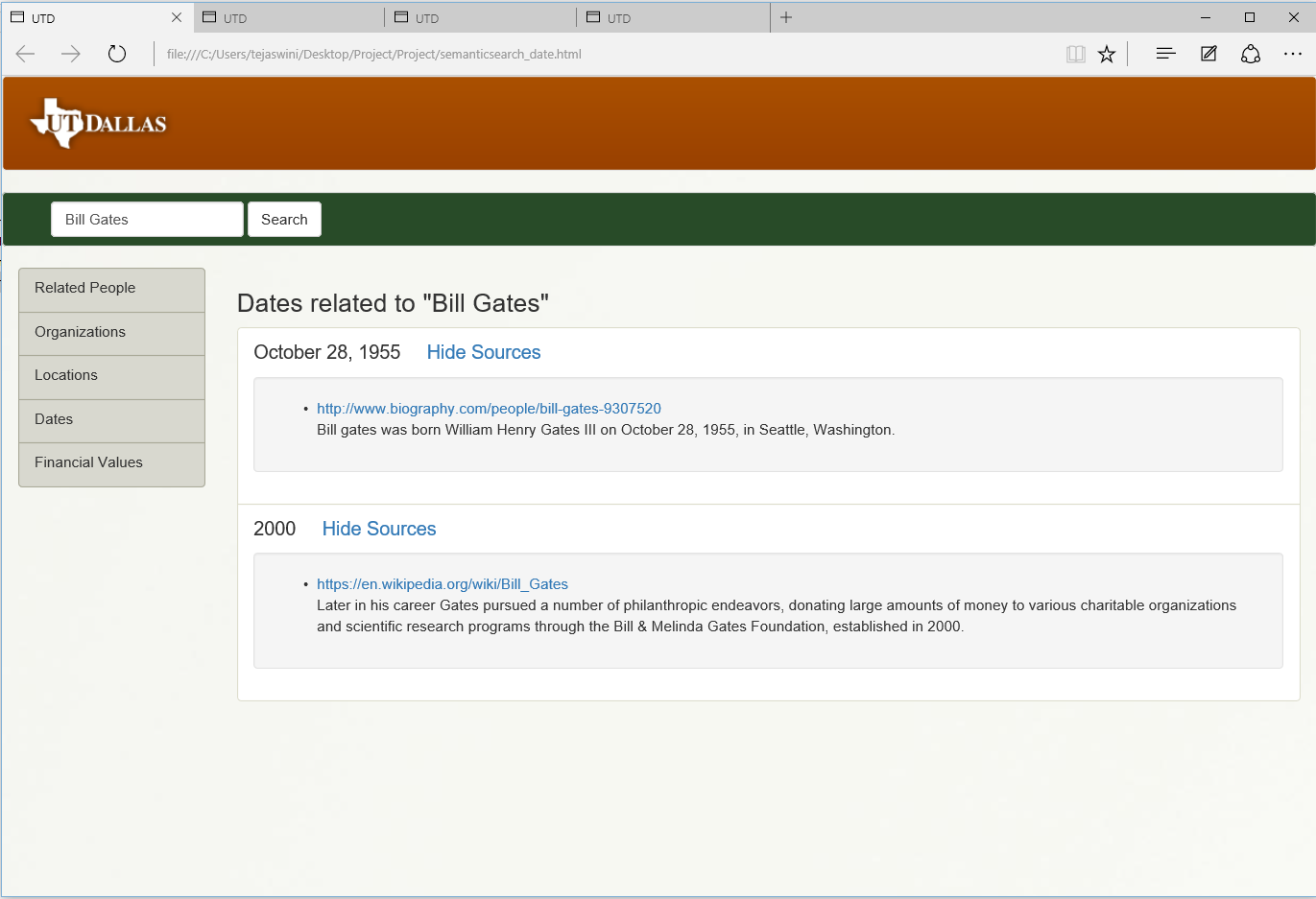


Figure 4: Person Related to Dates

The same entity related to other people is displayed as follows:-

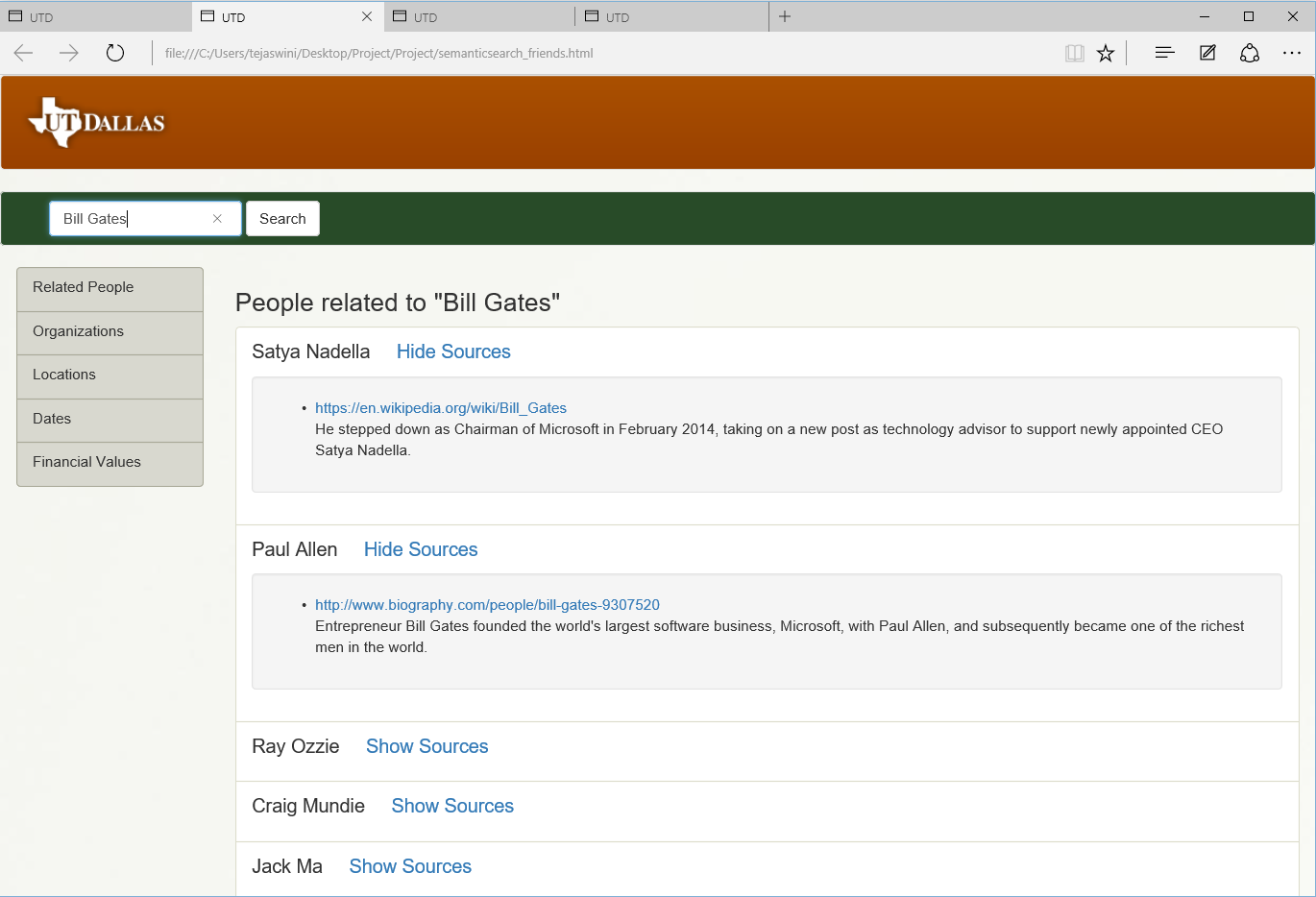


Figure 5: Same Example With People Relations

And location results are displayed as:-

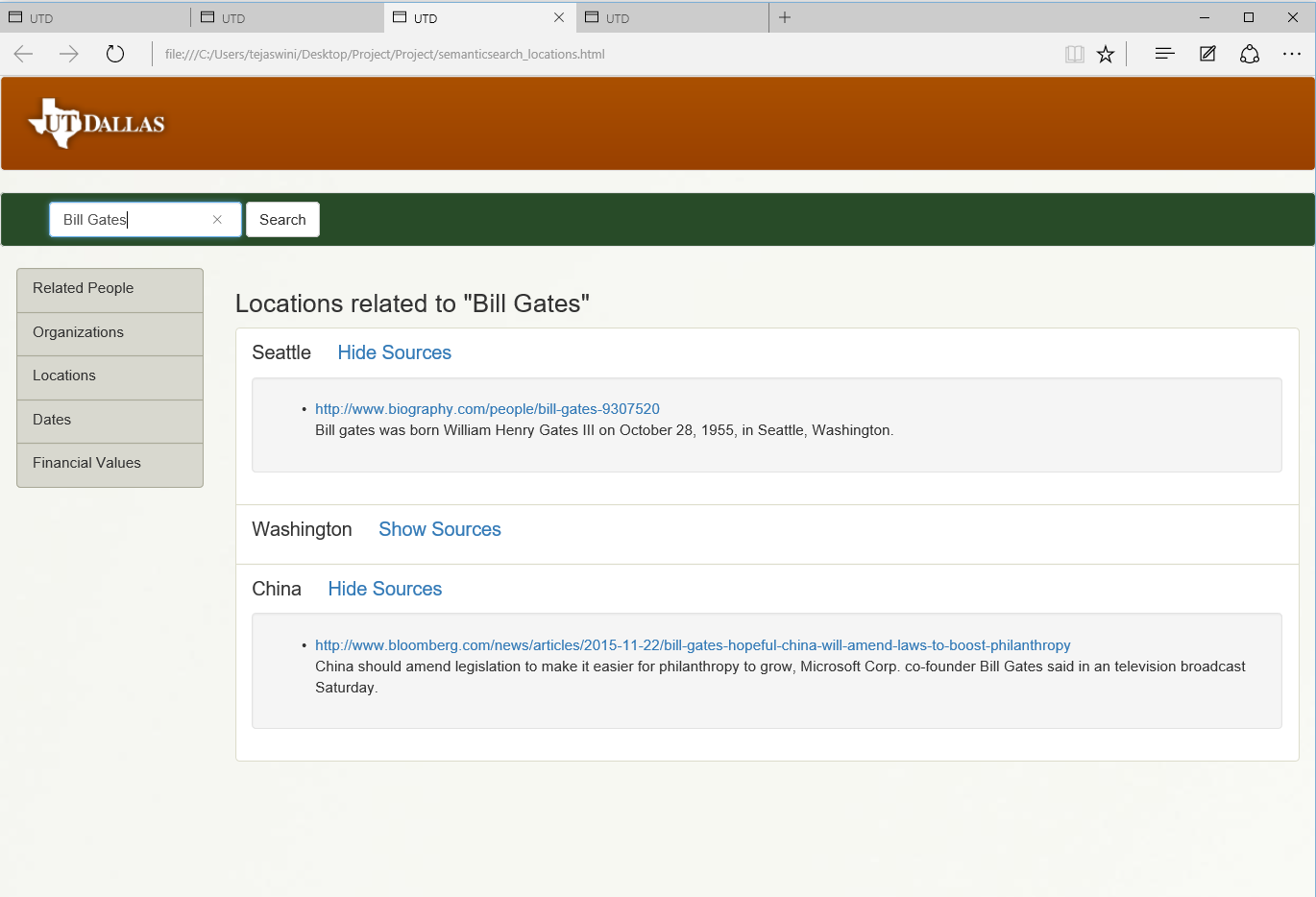


Figure 6: Location Results

And similarly, the organization linked to the person are displayed as follows:-

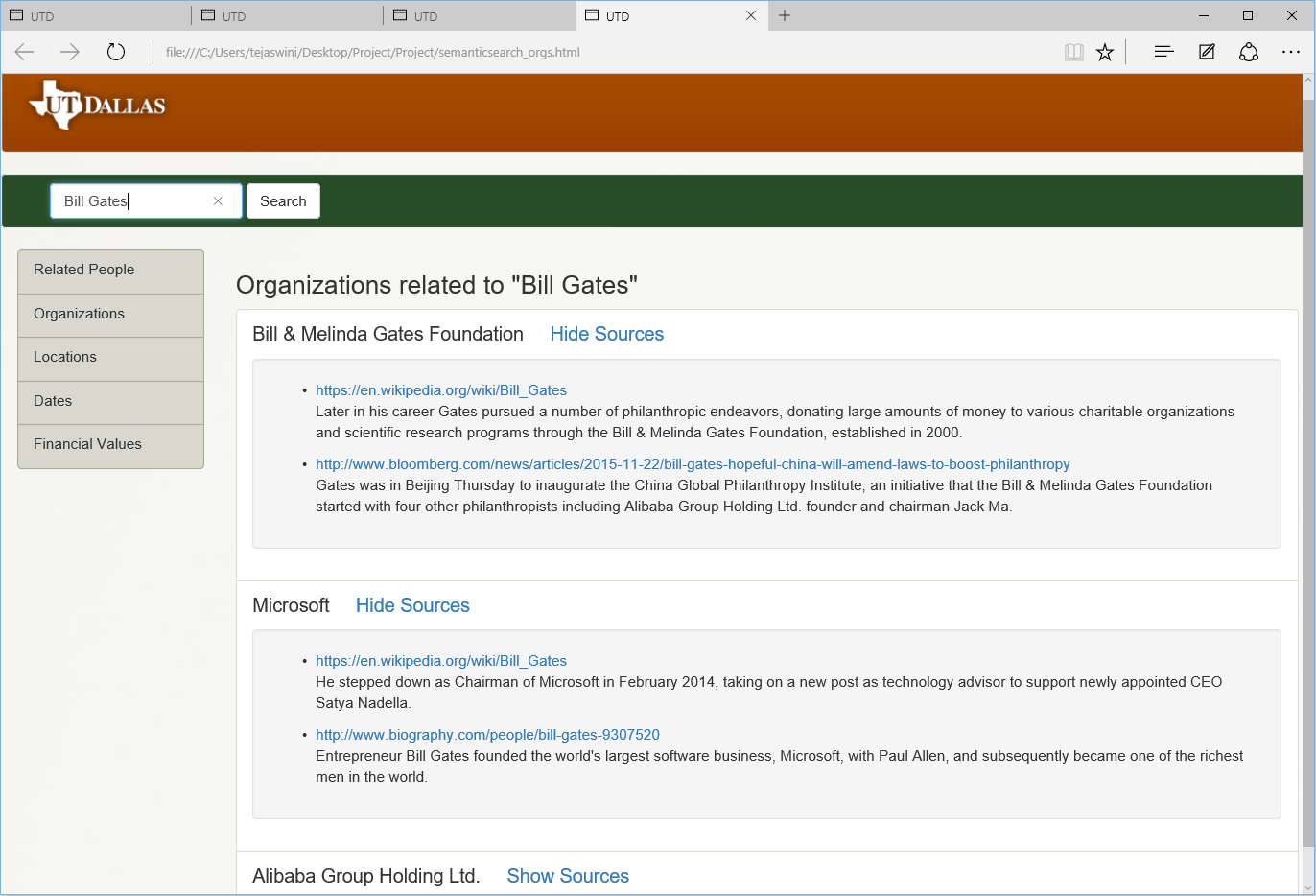


Figure 7: Related Organizations

## Custom project Justification

This project differs from an LOGD project in the following ways:-

1. Different source of data – The LOGD website was not used as the data source
2. Wider range of tools used – Including the use of Jena, Gruff, Protégé, other technologies like SolR, HTML, CSS, JavaScript, Java, Stanford API were also used for end-to-end integration
3. Custom Ontology – A custom ontology was defined to model and structure the data and the results displayed were structured using this model.

# Summary

As discussed, this project successfully extracts relations in unstructured data using Semantic Web technologies. As shown in the figures, these relations in the data and successfully displayed on a website for the user to derive meaningful information from.

The report also described the relevance of these results, the target audience for whom these result may prove to be relevant and the justification for this custom project. With the use of technologies as discussed in the report and the accuracy of the results obtained, this project can very well be useful for the audience it aims to target.

Although a small sample of data mainly concerning Bill Gates was used to discuss the results in this report, the project can be scaled to much larger size of data and can present to the user a much wider expanse of entities to search and derive structured information about.