Visual Analytics Ontology-guided I-DE system: A case study of Head and Neck Cancer in Australia

Ahsan Morshed[1], Abdur Rahim Mohammad Forkan[1], Tejal Shah[2], Prem Prakash Jayaraman[1], Dimitrios Georgakopoulos[1], Rajiv Ranjan[2]
[1]*Department of Computer Science and Software Engineering, Swinburne University of Technology, Melbourne, Australia*
[2]*School of Computing, Urban Sciences Building, Newcastle University,*
{amorshed, fforkan, pjayaraman, dgeorgakopoulos} @swin.edu.au & {raj.ranjan, tejal.shah}@ncl.ac.uk

***Abstract*— Visual analytics is inherently multi-disciplinary in nature encompassing diverse concepts from a wide range of scientific disciplines. It is a cutting-edge technology for multi-dimensional data analysis. In this paper, we discuss the development of a Visual Analytic Ontology (VAO) that provides a standard, unified, machine-understandable representation of expert knowledge that can be used to create rules to aid in the visual analysis of data. Further, we describe the development of Intelligent Data Ecosystem (I-DE) that uses the VAO to visualize the analytic information. We exemplify the efficacy of I-DE using the head and neck cancer case study data from Australia.**

Keywords*— *Visual analytics, ontology design, Intelligent Data Ecosystem, Head and Neck Cancer, Weather.

I. INTRODUCTION

Visual analytics is the semantic glue that connects the pieces of complex data to facilitate the effective exploration of data through meaningful visualizations [1]. The visual analytics literature is vast and provides us with a large body of knowledge about domain specific as well as more generic visual analytics techniques, frameworks, and studies. In addition to the major visual analytics journals and conferences, this include results and concepts from a range of knowledge generation and verification paradigms such as exploratory data analysis (EDA) [2] and knowledge-discovery in data mining (KDD) [3, 4]. The field of visual analytics has evolved from these paradigms by specifically emphasizing the visual and the human-in-the-loop components. However, while many of the underlying concepts remain the same, the number of names that have emerged over the years has become confusing, in particular for non-experts. Examples of knowledge-discovery paradigms that specifically employ visualization as part of the knowledge discovery process are: statistical graphics, exploratory data analysis, knowledge-discovery in data mining [3, 4], visual data analysis [5], visual data exploration [7], visual data mining [6, 7], and visual analytics [1]. In addition, it is an ongoing effort to relate some of these paradigms into a unified model of the knowledge generation process [8] but there is no structured approach to formalize the concepts used in visual analytics (and many of the other paradigms). Thus, there is a need to for a unified knowledge representation framework to support visual analytics.

Defining a unified framework to understand the multi-disciplinary nature of visual analytics and summarize the key relationships between the abundance of visual analytics concepts would facilitate the effective exploitation of the power of visual analytics. In order to understand the relation between these concepts (across disciplines, paradigms, and communities), enable automatic processing, detect ambiguous concepts (e.g. 'pattern' in KDD vs. 'pattern' in visualization), we develop an ontology for visual analytics by reusing relevant existing ontologies and also demonstrate a prototype Intelligent Data Ecosystem (I-DE) where the ontology is used to guide data analysis.

Rest of the paper is organised as follows. In Section II, we describe Visual Analytic Ontology (VAO) design and implementation, in Section III, we present the abstract level of the VAO and in Section IV, present an ontology guided application together with the visual analysis results. We conclude the paper in section V.

II. VISUAL ANALYTIC ONTOLOGY DESIGN AND IMPLEMENTATION

Ontology refers [9] to the formal specification of entities or objects and their relationships in a domain. Being machine-interpretable, ontologies can be used for automated processing tasks such as decision support. Further, ontologies contain domain information and hence can provide rich contextual knowledge for data analysis. Some work [10, 11, and 12] has been done on integrating ontologies into visual analytics platforms. In a recent European project called CODE (Commercial empowered Linked Open Data Ecosystems in Research) [10], ontologies are used for describing visual components. In visual analytics, ontologies play a vital role in providing appropriate meaning to data and ensure data quality. For example, a health record data containing ECG or Gene Expression data is not meaningful without proper metadata to give context to the data.

Figure 1 gives an overview of the VAO development process. As a first step, we will identify relevant concepts and relationships from literature, existing ontologies, and other data sources that would provide the requisite background knowledge and metadata for analysis. Next, we will implement the concepts and relationships in the Web Ontology Language (OWL), which is a W3C recommendation [13]. Based in Description Logic, OWL is an expressive ontology language with inferencing capabilities that is widely used to model rich and complex

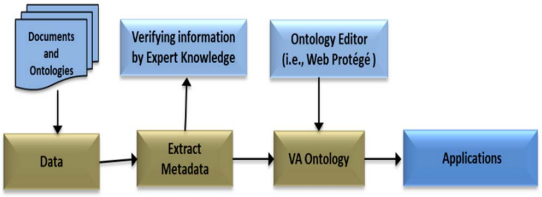


Fig. 1. A workflow of data to ontology

978-1-5386-9502-9/18/$31.00 ©2018 IEEE
DOI 10.1109/CIC.2018.00064

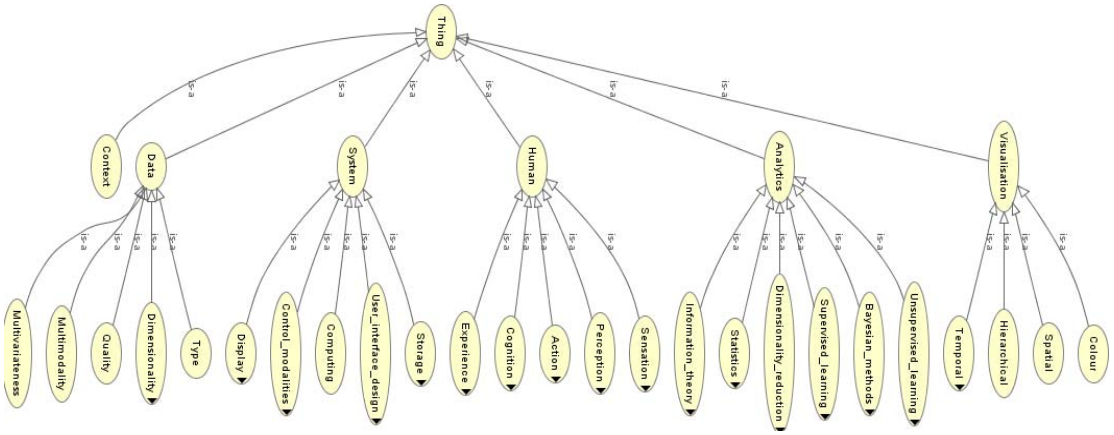


Fig. 2. Overview of VA ontology

ontologies. The VAO, implemented in OWL, will thus be able to provide the ground truth for supervised and unsupervised machine learning algorithms as well as act as a knowledgebase. Specifically, it will help to disambiguate the concept meaning from a given text in the VA domain. A human brain can classify various concepts easily when it possesses the requisite background knowledge. For example, consider the word "Java" that has three core meanings [14] being i) an island in Indonesia ii) a programming language iii) a beverage consisting of an infusion of ground coffee beans. When used in a sentence, a human brain can understand the context and hence determine which "java" is being referred to. Similarly, a machine would require an appropriate dictionary to determine the intended meaning. The use of word sense disambiguation techniques, with the VAO acting as a knowledge base (dictionary), will thus allow the most appropriate meaning to be determined for provided query words. Furthermore, VAO act as an enabler to guide a system to find different patterns from data and provide the appropriate meaning.

III. Visual Analytics Ontology

The VAO has been modelled using the open source Protégé ontology editor [15]. It allows human experts to capture their knowledge in formal way and later captured knowledge can be converted to into rules. These rules can help to machine learning algorithms for the decision making process. Figure 2 shows the overview of the development process of VAO and Figure 3 presents the hierarchical layout of the ontology. VAO has six abstract level of concepts such as Context, Data, System, Human, Analytic and Visualisation. The details are as follows:

Context: this concept defines the situation of cases such as "Head Neck Cancer"

Data: This concept contains all the information about the curated data from heterogenous data sources. It has sub concepts dimensionality, multimodality, multivariances and types. Head Neck Cancer and Weather data for corresponding time period exemplify heterogenous data source.

System: This concept illustrates computing, display and user interface design concepts as part of system concept. This concept defines how the prototype will be look like. For example, I-DE tool.

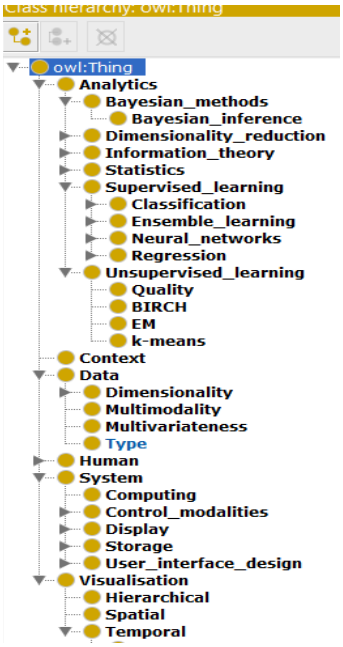


Fig. 3. Ontology layout from protégé tool

Human: This concept captures human activities, cognitive functions demographic information and experiences. Such information helps to model domain information in a structured way.

Analytics: This is concept is very important part of VAO. It contains all information about the analytics methods such as supervised, unsupervised and statistical. This concept helps in cleaning, pre-processing and analysis of the curated data. For example, we have used correlation for analysis of the head neck cancer and weather data.

Visualisation: This abstract concept has hierarchical, spatial and temporal sub concepts as part of the visualisation concept. This concept helps to visualise the information either hierarchical or spatio-temporal way. For example, the developed I-DE tool visualise information in spatio-temporal and hierarchical way.

The proposed VAO is incorporated in our prototype Intelligent Data Ecosystem (I-DE) tool. In the I-DE, VAO is extended with association rules for improved reasoning and querying over the data as well as evaluating the quality of metadata. In the next section, we demonstrate the use of VAO in the I-DE with real-world neck and cancer data in Australia.

IV. A CASE STUDY OF HEAD AND NECK CANCER TO FOR VISUAL DATA ANALYSIS

We evaluated the proposed VAO to visualize the statistics of head and neck cancers in a specific geographical region. The original data source contains a global view of all cancer types. However, we demonstrate how the VAO can be used as a guide to interpret the distribution of a specific cancer type – in this case, head and neck cancers. The rate of incidents for head and neck cancer types varies in different geographical regions. Accordingly, we also demonstrate how a correlation can be visualised across multiple data sources. In order to do this, we use climate data along with the cancer data to identify the correlation between a specific cancer trend with the weather conditions in a given region.

A. About Head and Neck Cancers

Head and neck cancers are amongst the ten most common cancers in both men and women and the seventh most commonly diagnosed cancer types in Australia [16]. Head and neck cancers constitute a group of cancers that usually begin in the squamous cells that line the insides of the mouth, nose and throat. In 2014, there were 4537 incidents of these cancers diagnosed in Australia (3,342 males and 1,195 females). In 2018, it is estimated that 5,091 new cases of head and neck cancers will be diagnosed [16].

B. Data collection and caetrgorisation

As our data source we have used cancer data collected by the Australian Institute of Health and Welfare and made available publicly [17]. This data contains the annual total counts of all cancer incidents in Australia and mortality by sex, year, age and type (from 1980 to 2016). As our metadata, we extracted incidents of only head and neck cancers from this data source. We divided this data into five categories based on incident counts for five years (2012 to 2016).

Further, we extracted climate data (1990 to 2016) available from the Bureau of Meteorology website [18]. This set contained monthly weather statistics such as rainfall, maximum temperature, solar exposure, and mean of maximum temperature near a weather station of a particular region. From this, we separated the data of interest – weather data in our region of interest – for analysis. The collected weather data was then matched with the yearly statistics of cancer data. Few metadata were missing, and some metadata were illustrated in different semantic, for example, temperature often described in Celsius or Fahrenheit. To tackle these challenges, the VAO has been used to check the quality and classify the information as required by the I-DE tool.

V. I-DE FOR VISUAL ANALYTICS

To demonstrate the advantages of visual analytics and the proposed VAO, we have developed a tool called I-DE or Intelligent Data Ecosystem to visualize information related to head and neck cancers in Australia in different years via an interactive dashboard. A sample view of the I-DE is presented in Figure 4.

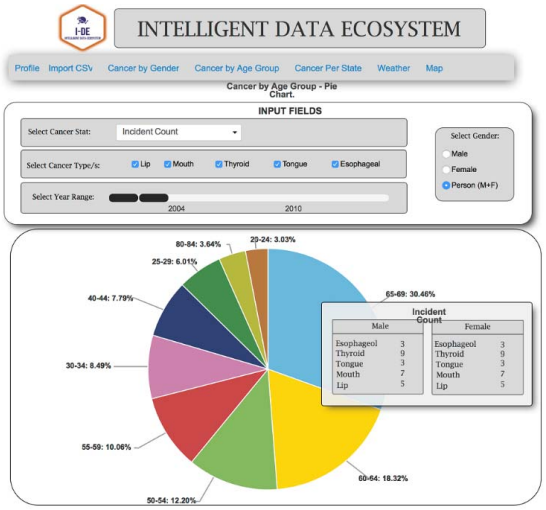


Fig. 4. A view of our developed visual analytics tool – Intelligent Data-Ecosystem

We have integrated open heat maps [19] to display the statistics of our targeted case study (head and neck cancer incidents) across Australia using map based visualization. The open heat map requires some parameters for generating colored heat map such as geographical location, category, color and numerical counts.

Figure 5 displays the areas of head and neck cancer incidents across Australia as well as in individual states (highlighted separately). The high incident areas are represented by darker colors. This map shows that head and neck cancer trends are different across the country. We have

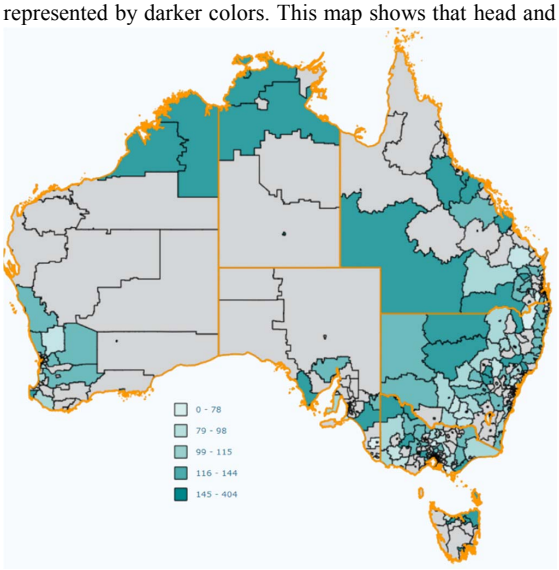


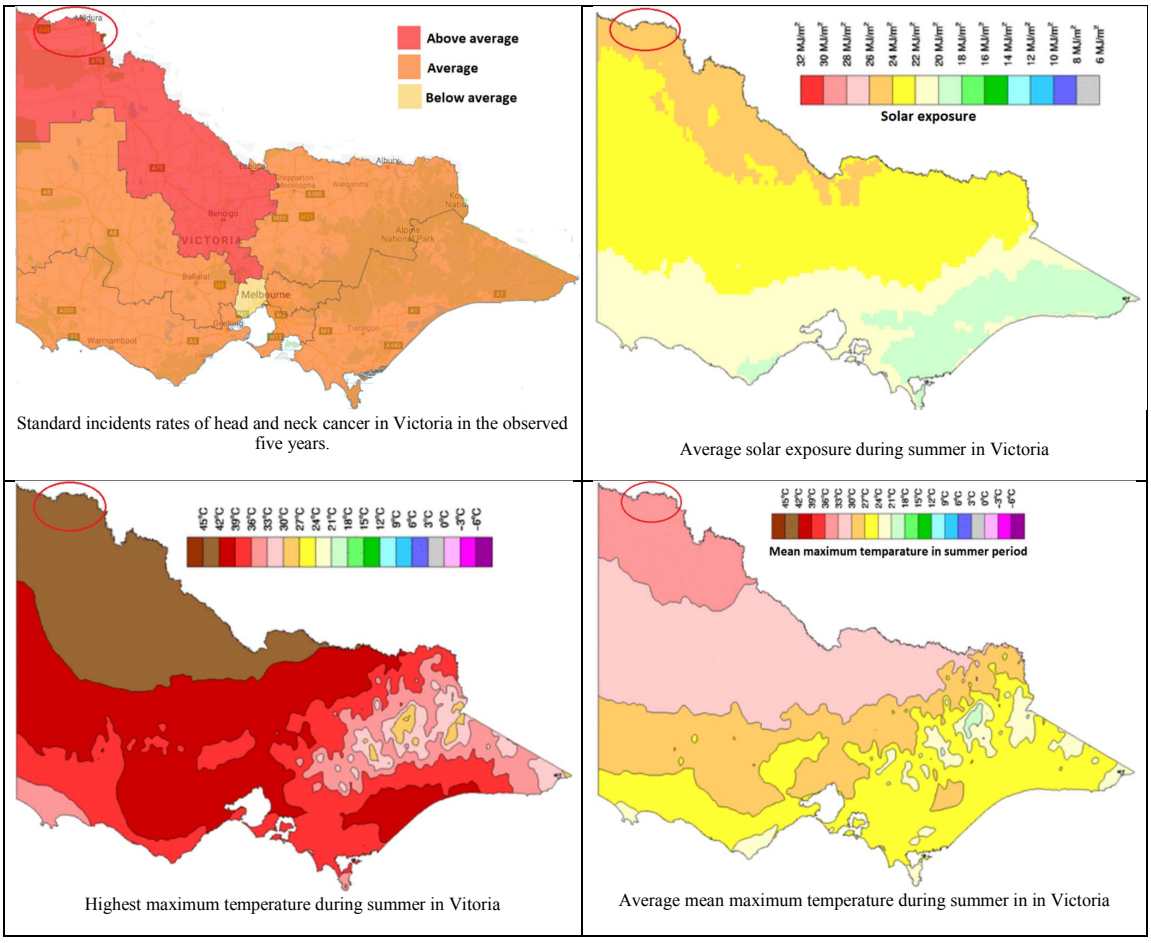
Fig. 5. Statistics of Head and Neck cancer across Australia in 5 years

used Victoria as our case study area for further analysis using the VAO.

In order to visualize the intensity of cancer incidents across different areas of Victoria, we further classified the incident rate into three categories (average, below average and above average). Following that, we used the interactive map view to visualize information about head and neck cancer incidents as well as summer climate condition (weather data) over a period of five years in Victoria (see Figure 6). The intensity represents the cancer incidences per 100,000 population in each region.

From the interactive heat maps in Figure 6, we can observe that the rate of incidents of head and neck cancers is higher in north-western side (highlighted using red circles) of Victoria where we can also see high solar exposure (i.e. UV index), mostly high summer temperature and high average mean temperature during the summer period. A closer view in terms of number of incidents in observed five years (Figure 7) also confirms that the incidents are high in areas with extreme climate conditions during the summer period.

The visual analytics developed from our ontology shows a high degree of correlation in the intensity of cancer incidents and weather conditions. Similar correlation can also be observed when the yearly changes to maximum temperature are visualized with the number of incidents using our developed I-DE. This is shown in Figure 8.



Standard incidents rates of head and neck cancer in Victoria in the observed five years.

Average solar exposure during summer in Victoria

Highest maximum temperature during summer in Vitoria

Average mean maximum temperature during summer in in Victoria

Fig. 6. Head and Neck cancer density, and some feature distribution in Victoria in the observed 5 years

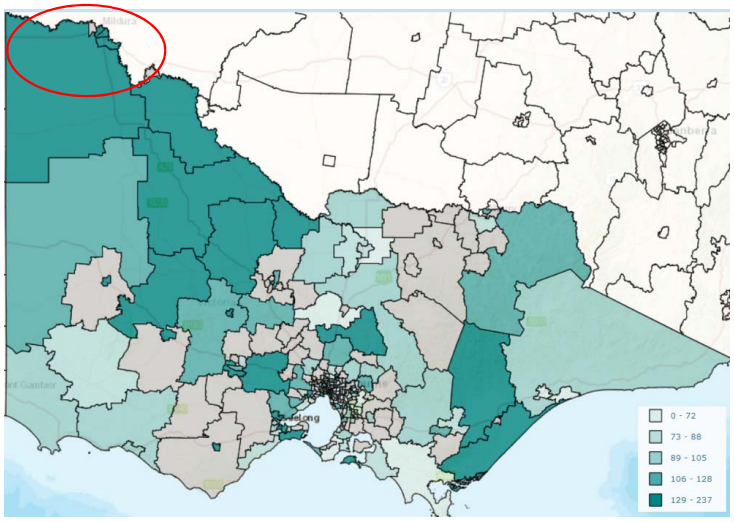


Fig. 7. Head and Neck cancer incidents in different region of Victoria in obserevd 5 years.

The top plot of Figure 8 shows the trends of head and neck cancers history over 27 years (from 1990). The weather data of high incident area (Mildura) has been explored and integrated in our visual analytics with this trend. The bottom plot of Figure 8 shows the highest maximum summer temperature across 27 years for Mildura region. We identified a high correlation between this extreme weather and the rate of head and neck cancer incidents in different years. The incident rate mostly increases or decreases with the rise and fall in maximum temperature.

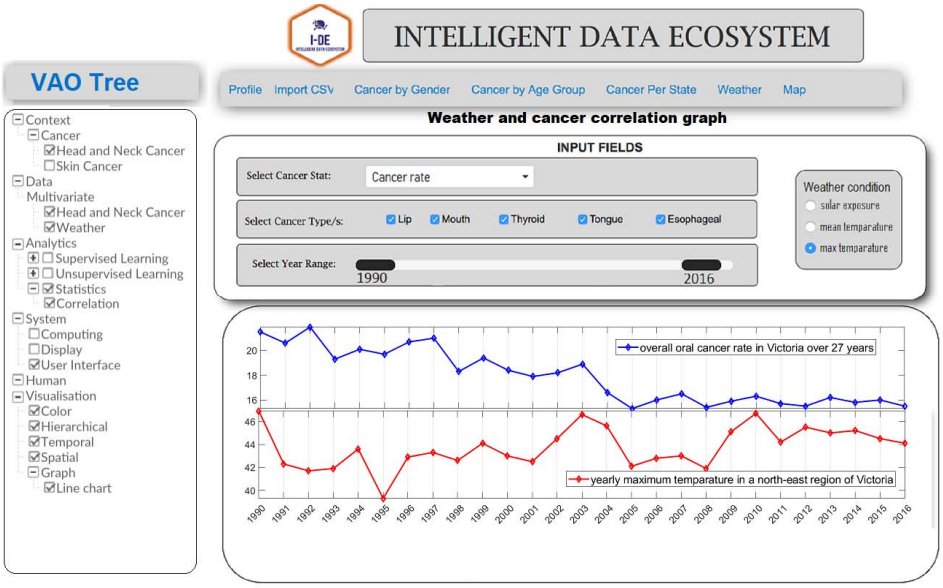


Fig. 8. A relation with yearly highest maximum temperature with head and neck cancer incident rate in the cancer-prone regions of Victoria.

VI. Conclusions

We presented the development of a Visual Analytics Ontology and developed an ontology-guided visual data analytic tool called I-DE. I-DE provides different interactive visualization capabilities for visualizing different patterns and trends in the data. We used an example of head and neck cancer data in Australia to demonstrate the efficacy of the proposed ontology and the I-DE tool. As a next step, we plan to extend this work by adding more datasets and investigating increasingly complex patterns from the data.

REFERENCES

[1] J. J. Thomas and K. A. Cook (Ed.), "*Illuminating the Path: The R&D Agenda for Visual Analytics,"* National Visualization and Analytics Center, p. 4, 2005.

[2] J. W. Tukey, *Exploratory Data Analysis*. Massachusetts California London Amsterdam Ontario Sydney: Addison-Wesley Publishing Company, 1977.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[5] D. Keim, "Information Visualization and Visual Data Mining," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[6] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in Visual Data Analysis," in *Proceedings of the Tenth International Conference on Information Visualisation*, 2006, pp. 9–16.

[7] M. C. Ferreira de Oliveira and H. Levkowitz, "From Visual Data Exploration to Visual Data Mining: A Survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 3, pp. 378–394, 2003.

[8] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge Generation Model for Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604–1613, Dec. 2014.

[9] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

[10] European Project. See at http://code-research.eu/visual-analytics/

[11] Wong, Pak Chung, George Chin, Harlan Foote, Patrick Mackey, and Jim Thomas. "Have Green-a visual analytics framework for large semantic graphs." In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pp. 67-74. IEEE, 2006.

[12] Vehlow, C., Kao, D. P., Bristow, M. R., Hunter, L. E., Weiskopf, D., & Görg, C. (2015). Visual analysis of biological data-knowledge networks. *BMC bioinformatics*, *16*(1), 135.

[13] W3C Working Group, "OWL 2 Web Ontology Language Document Overview," W3C, 2012. https://www.w3.org/TR/owl2-overview/

[14] WordNet. See at http://wordnetweb.princeton.edu/perl/webwn

[15] Protégé . See at http://protege.stanford.edu/

[16] Head and Neck Cancer statistics. See at https://head-neck-cancer.canceraustralia.gov.au/statistics

[17] Australian public data repository: See at https://search.data.gov.au/

[18] Bureau of Meteorology: See at http://www.bom.gov.au/climate/data/

[19] Open Heat Map: http://www.openheatmap.com/