

# PrEVIEW: Clustering and Visualising PubMed using Visual Interface

Syeda Sana e Zainab<sup>1</sup>, Qaiser Mehmood<sup>1</sup>, Durre Zehra<sup>1</sup>, Dietrich Rebbholz-Schuhmann<sup>1</sup>, and Ali Hasnain<sup>1</sup>

Insight Centre for Data Analytics, National University of Ireland, Galway  
firstname.lastname@insight-centre.org

**Abstract.** The life sciences domain has been one of the early adopters of Open Data Initiative and a considerable portion of the Linked Open Data cloud is comprised of datasets from Life Sciences Linked Open Data (LSLOD). This deluge of biomedical data and active research over the past decade resulted in the flux of scientific publications in this domain. PubMed resource provides access to MEDLINE, NLM's database of citations and abstracts in the biomedical domain. PubMed Central provides links to full-text articles along with publisher web sites, and other related resources. In this paper we present PubMed Visual Interface (PrEVIEW)- a web based application to access information related to *publication, research topic, author* and *institute* through a visual interface. PrEVIEW not only provides useful information e.g. research topic of interest, research collaboration at personal or institute level etc, for the biomedical research community but also helpful for the working Data Scientist. We also evaluate the usability of our system by using the standard system usability scale as well as a custom questionnaire, particularly designed to test the usability of the interface. Our overall usability score of **83.69** suggests that web based interface is easy to learn, consistent, and adequate for frequent use.

**Keywords:** PubMed, Publication, Visual Interface

## 1 Introduction

The deluge of biomedical data in the last few years, partially caused by the advent of high-throughput gene sequencing technologies [3,5,2], has been a primary motivation for efforts in this area. The significant contributors includes the Bio2RDF project<sup>1</sup>, Linked Life Data<sup>2</sup>, Neurocommons<sup>3</sup>, Health care and Life Sciences knowledge base<sup>4</sup> (HCLS Kb) and the W3C HCLSIG Linking Open Drug Data (LODD) effort<sup>5</sup>. These efforts have been partially derived and are

<sup>1</sup> <http://bio2rdf.org/> (l.a.: 2016-03-31 )

<sup>2</sup> <http://linkedlifedata.com/> (l.a.: 2016-03-01 )

<sup>3</sup> [http://neurocommons.org/page/Main\\_Page](http://neurocommons.org/page/Main_Page) (l.a.: 2016-03-16 )

<sup>4</sup> <http://www.w3.org/TR/hcls-kb/> (l.a.: 2016-03-16 )

<sup>5</sup> <http://www.w3.org/wiki/HCLSIG/LODD> (l.a.: 2016-07-16 )

still motivated by the deluge of data in biomedical facilities in the past few years, partially caused by the decrease in price for acquiring large datasets such as genomic sequences (e.g. the cost of sequencing the genome has dropped faster than Moore's law) and the trend towards personalised medicine, pharmacogenomics and integrative bioinformatics. This increase in biomedical research has resulted in the drastic increase in scientific publications in this area. This includes both Conference as well as Journal publications. In order to investigate the ongoing research trends, similar or related research contribution and possible research collaboration one has to keep an eye on the research papers from peer scientist. There is a need for an integrated system where one can find the answers for the questions e.g. *List of all the authors who publish their research on Lungs Cancer in any conference*. With the emergence of PubMed it become possible to access the MEDLINE, NLM's database of citations and abstracts in the biomedical domain including medicine, nursing, veterinary medicine, dentistry, health care systems, and preclinical sciences. PubMed Central provides links to full-text articles along with publisher web sites, and other related resources. Using PubMed, one has to access their web interface in order to search for any publication or through RESTful Web Services. Using these RESTful services poses limitations for Biomedical researchers searching for relevant articles as they are domain user with limited or no knowledge of using such services[15,4]. In this paper we introduce PrEVIEW a web based application that uses RESTful Web Services provided by PubMed and user can search for any topic e.g. *Cancer*, and all the relevant information regarding the publications about the searched concepts along with the publication type, author and institutes involved in the *Cancer* research are retrieved through graphical and intuitive interface. This interface make it easier to cluster Authors working on similar topics or institutes involved in any particular research area.

The remaining part of this paper is organised as follows: we highlight the related work in section 2. We introduce our methodology and PrEVIEW salient features in section 3. Later we present the usage scenario in section 4. Subsequently, we present an evaluation of our approach in section 5. We finally conclude the paper.

## 2 Related Work

AMiner by Tang J et. al creates semantic-based profile for each researcher in order to built researcher social network.[12] Osborne F Et. al introduce tool Rexplore which mainly focus on relating authors semantically in order to understand the dynamics of research area. [11] Monaghan F et. al introduce application Saffron which extract information from unstructured documents using Natural Language Processing techniques.[10] Wu et. al [13] visualised author collaboration network for schizophrenia disease, between year 2003 to 2012 using CiteSpace III visualisation. Xuan et. al [14], developed Medline exploration approach for interactive visualisation. They made use of PubViz, for grouping, summary, sorting and active external content retrieval functions. Similarly Joseph T et. al, introduced

TPX [7] which is a web-based PubMed search enhancement tool that enables faster article searching using analysis and exploration features in tabular form. In the book “Analysis and Visualisation of Citation Networks” Zhao et. al [16], discussed mapping research fields through citation analysis. H Chen et. al presented NLP-based text-mining approach, “Chilibot”, [1] which constructs content-rich relationship graph networks among biological concepts, genes, proteins, or drugs from pubmed (abstract). Table 1 shows a brief comparison of PrEVIEW with similar systems.

Table 1: Comparison of Applications based on Pubmed Data (Time Limit **TL**, Topic Limit **TLs**, Visualisation **Vs**, Explorative Search **ES**, Topic Structuring **TS**, Time Based Analysis **TbA**, Community Structure **CS**, Social Network Focused **SNF**, Genes Proteins Drugs Relations **GPDR**, Schizophrenia Disease Only for year (2003-2012), **SDO**, Tabular Output Only **TOO**, Most Cited Only **MCO**, Mediline Only **MO**)

Publications	Tool	TL	TLs	Vs	ES	TS	TbA	CS	Limitations
Tang J et. al [12]	AMiner	X	X	X	✓	✓	✓	✓	SNF
Osborne F Et. al[11]	Rexplore	X	X	X	✓	✓	✓	✓	SNF
Monaghan F et. al[10]	Saffron	X	X	X	✓	✓	✓	✓	SNF
H Chen et. al [1]	Chilibot	X	✓	✓	X	X	X	X	GPDR
Ying Wu et. al [13]	CiteSpace III	✓	✓	✓	X	✓	X	✓	SDO
Joseph T et. al[7]	TPX	X	X	X	✓	✓	X	X	TOO
Zhao et. al [16]	Citation Networks	X	X	X	X	X	X	X	MCO
Weijian Xuan et. al [14]	PubViz	X	X	✓	X	✓	X	✓	MO

The table 1 shows that PrEVIEW is not limited for searching particular topic, not confined for searching articles for specific time frame, or specific database of pubmed neither for the social networking purposes. Furthermore tabular display, can be difficult to pull off linked information while concept map approach [6] used by PrEVIEW show better interactions. PrEVIEW can also retrieved results based on most recent publications instead of most cited ones [16]. PrEVIEW also supports searching and navigation through visualisation that makes it more intuitive to use.

### 3 Methodology

Our methodology consists of two steps namely: 1) result retrieval using Rest API’s and 2) building visual interface .

**Result Retrieval Using Rest API’s** REST (Representational State Transfer) is a communication approach that is often used in the development of Web services<sup>6</sup> and involves reading any Web page that contains an XML file. PrEVIEW

<sup>6</sup> <http://searchsoa.techtarget.com/definition/REST>

make use of REST APIs<sup>7</sup> provided by PubMed and overall result retrieval works as follow: (1) User search a *topic, institute, title of the publication or author* and send the request, (2) Rest API's fetch data in the form of XML, (3) XML parser with data manipulator creates JSON file and transfer to presentation layer.

**Building Visual Interface** The analysis of large numbers of biomedical publications, in the form of graphical representation made it possible for researchers to extract the relevant publications for future collaboration. We chose the concept map approach [6] for building the visual interface, which is a graphical method representing the relationship between nodes and links, and has been used in various domains for organising knowledge. Using this approach we represent concepts of title, topic, authors and institutes and relevant publications. Greater the occurrence of any searched concept bigger the size of the circular node representing that concept. Due to large amount of data we only present top 100 cited occurrence for any searched concept. Subsequent results are displayed in a tabular form using "View All Records" option. Interface also provides the graphical representation of topic, authors or institute and respective occurrence (the number of times it occur in the result).

**Technologies** PrEVIEW is browser-based client application that provides a flexible front-end for searching. To build this application variety of web technologies are used including HTML5, CSS, JavaScript, JQuery<sup>8</sup>, Rest API's<sup>9</sup>, Java Servlet, SVG<sup>10</sup>, AJAX<sup>11</sup> and JSON<sup>12</sup>.

**Availability** The PrEVIEW application can be accessed at <http://srvga178.derri.ie/PrEVIEW/>.

### 3.1 System Overview

Figure 1, presents the architecture of PrEVIEW that comprises of three layers namely "Data Layer", "Business Logic Layer" and "Presentation Layer". Mining PubMed data brings two challenges. The first challenge is to retrieve data from PubMed, based on their terminological, author or institute content. The second challenge is to classify and visualise the results on the basis of selected field. For addressing the first challenge, at the Data Layer, we use Rest APIs, provided by PubMed to retrieve user searched term directly from PubMed resource in the form of XML. The business layer parses XML data and create respective JSON file. For example in case of topic search "cancer", XML of

<sup>7</sup> <http://europepmc.org/restfulwebservice>

<sup>8</sup> <https://jquery.com/>

<sup>9</sup> <http://searchsoa.techtarget.com/definition/REST>

<sup>10</sup> [www.w3schools.com/svg/](http://www.w3schools.com/svg/)

<sup>11</sup> <http://api.jquery.com/jquery.ajax/>

<sup>12</sup> <http://json.org/>

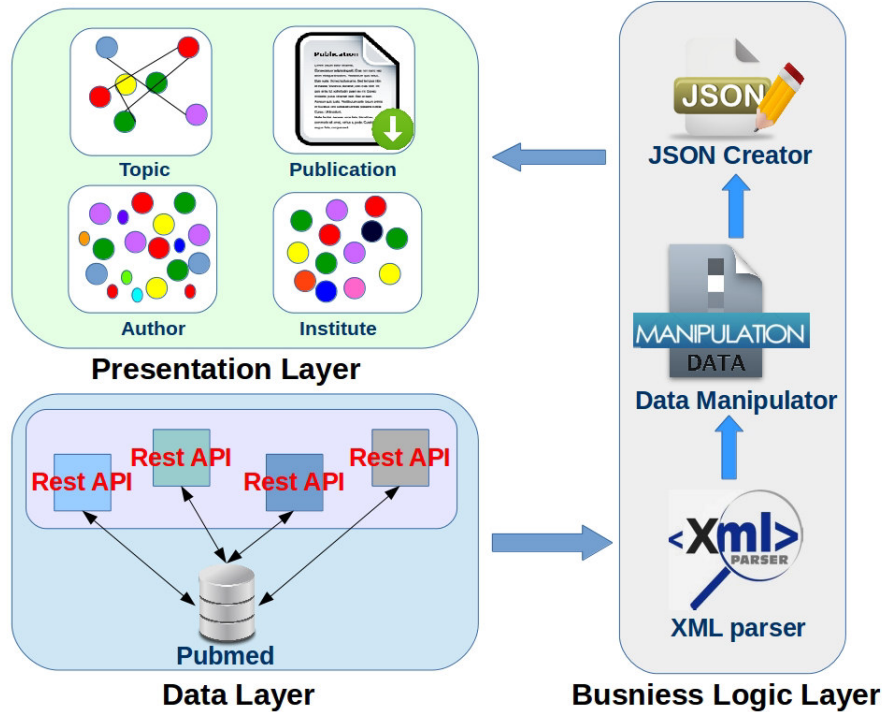


Fig. 1: PrEVIEW Architecture: Data, Business and Presentation layer.

all the relevant publications is retrieved at the *Data Layer* and corresponding JSON is generated at the *Business Layer*. At the *Presentation Layer*, graphical and visual representation of PID's (PubMed Id's), authors and institute are displayed as clusters. Complete metadata description of publications can be accessible by clicking on PID's . For the second challenge, the classification, we adopt the technique that relies on publication mapping, a web-based client application PrEVIEW is developed. Publications are mapped in the form of clusters classifying author, topic, title and institute separately. Authors and institutes with remarkable contribution in regard to particular topic search, represented as larger node. Additionally, the publications can be downloaded. We applied two mapping techniques to this application: force directed placement [8] for clusters and chart summary for highly ranked search results.

#### 4 Usage Scenarios

Recognising the value of research networking, two example scenarios are discussed to demonstrate the use of PrEVIEW. First use-case visualises PubMed ID's and authors as the outcome from topic/title query. For second use-case,



Fig. 2: Highly cited publications with the author list on "cancer" research.

the institutional outcomes have been used. The combination of these two use-case enables establishing research networks and collaborations at authors and institution level. Details of these two outcomes have been discussed below.

**Highly cited publications with the author list having paper on "cancer" research** The step-by-step approach is discussed as follow (Fig. 2):

1. The first step is to specify the selection of the corresponding text. User search "cancer" while selecting Topic from the drop-down menu. (window A).
2. The visualisation of top 100 highly cited papers on cancer can be seen in (window B) where user can explore them using PubMed Id's and and description about the full paper can be explored (window D).



Fig. 3: research Institutes working on "cancer", and their collaborations with other institutes in the same domain.

3. List of all authors (from top 100 cited papers) working on cancer domain is shown in (window C) where a mouse-hover further display the "author's name".
4. Selected author contribution in cancer research (window E) provide user the ability to download all the relevant papers in cancer domain.

Authors collaborations at international level increase citations of research manuscripts. One of the novel advancement in the mode represented in Figure 2 is establishing authors research networks and collaborations on a focused domain.

**Research Institutes working on "cancer research", and their collaborations with other institutes** The step-by-step approach is discussed as follows(Fig. 3):

1. The first step is to specify the selection with the corresponding text. In this case "cancer" while selecting "Topic" from "drop-down menu" (window A).
2. The visualisation of highly cited cancer papers can be seen (window B) where user can explore them using PubMed Id's.

3. List of all institutes working on "cancer research" are shown in (window C) where user is able to explore any of them by selection.
4. Selected institute collaboration with other institutes in cancer research is shown in publication metadata, (window D).

The exponential growth in international collaboration on focused scientific research questions shows the novelty in the mode represented in Figure 3. Institutional collaboration enables shared learning, new research opportunities, establishing new research projects, joint applications for funds, and technology transfer. Findings from these different kinds of networks can be used in many ways. Collaboration networks in terms of highly cited authors, institutions, and countries are highlighted in the publication set. Semantic networks can identify various research directions the focused concepts. Further more publication citation networks can be used to quantify the citation impact in research directions and disciplines.

## 5 Evaluation

The goal of our evaluation is to quantify the usability and usefulness of PrEVIEW graphical interface. We evaluate the usability of the interface by using the standard *System Usability Scale* (SUS) [9] as well as a customised questionnaire designed for the users of our system. In the following, we explain the survey outcomes.

### 5.1 System Usability Scale Survey

In this section, we explain the SUS questionnaire<sup>13</sup> results. This survey is more general and applicable to any system to measure the usability. The SUS is a simple, low-cost, reliable 10 item scale that can be used for global assessments of systems usability[9].

As of 3rd May 2016, 22 users<sup>14</sup> including researchers and engineers in Semantic Web participated in survey. According to SUS, we achieved a mean usability score of **83.69** indicating a high level of usability according to the SUS score. The average scores (out of 5) for each survey question along with standard deviation is shown in Figure 4.

The responses to question 1 (average score to question 4.2 =  $\pm 0.72$ ) suggests that PrEVIEW is adequate for frequent use. The responses to question 3 indicates that PrEVIEW is easy to use (average score 4.5  $\pm 0.51$ ) and the responses to question 7 (average score 4.4  $\pm 0.59$ ) suggests that most people would learn to use this system very quickly. However, the slightly higher standard deviation to question 9 (standard deviation =  $\pm 0.74$ ) and question 10 (standard deviation

---

<sup>13</sup> SUS survey can found at: <https://goo.gl/hkuMDM>

<sup>14</sup> Users from AKSW, University of Leipzig and INSIGHT Centre, National University of Ireland, Galway. Summary of the responses can be found at: <https://goo.gl/iKaZjQ>



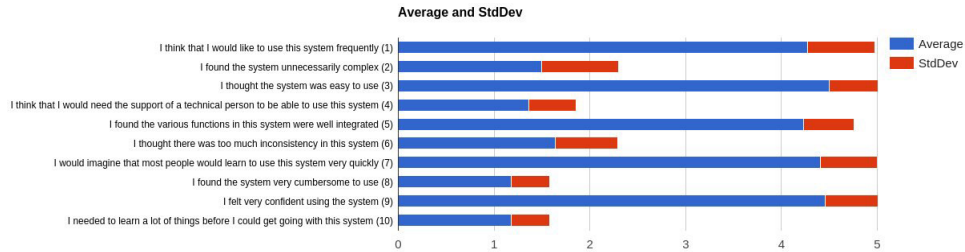


Fig. 4: SUS Evaluation

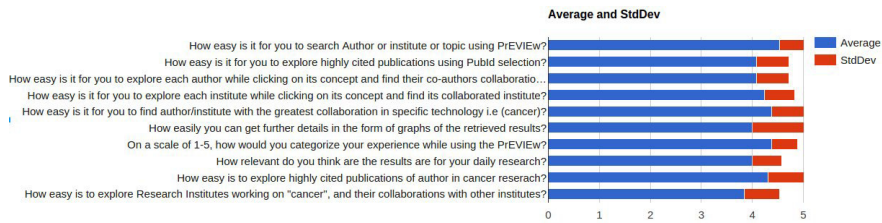


Fig. 5: Usability Evaluation

$\pm 0.39$ ) suggest that we may need a user manual to explain the different functionalists provided by the PrEVIEW interface.

## 5.2 Custom Survey

This survey<sup>15</sup> was particularly designed to measure the usability and usefulness of the different functionalists provided by PrEVIEW. In particular, we asked users to use our system and share their experience through question 9 and question 10. As of 3rd May 2016, 13 researchers including Computer Scientist<sup>16</sup> and Bioinformaticians have participated in survey. The average scores (out of 5 with 1 meaning strongly disagree and 5 meaning strongly agree) for each survey question along with standard deviation is shown in Figure 5. The average scores to question 10 (i.e.,  $4.45 \pm 0.73$ ), show that most of the users feel confidence to use the system and need not to learn much about the PrEVIEW before using. The responses to question 2 (average score =  $4.07 \pm 0.64$ ) suggest that exploring highly cited publications using PubId selection is easy in PrEVIEW. A slightly lower scores to question 7 (average score =  $4.38 \pm 0.51$ ) suggests that we need to further improve the user experience with visualisation components of the PrEVIEW.

<sup>15</sup> Custom survey can be found at: <https://goo.gl/syZzAM>

<sup>16</sup> Summary of the responses can be found at: <https://goo.gl/JKogDQ>

## 6 Conclusion and Future Work

In this paper we introduce PrVIEW as an online visual and graphical interface for searching PubMed resource. We evaluate our approach and usability of our system using the standard system usability scale as well as through domain experts. Our preliminary analysis and evaluation reveals the overall usability score of *83.69*, concluding PrVIEW an interface, easy to learn and help users accessing PubMed resource intuitively. As a future work we aim to extend PrVIEW with Faceted browsing and also provide visualisation at entity level e.g, Genes and Molecules where the search criteria retrieve these entities. Current work visualise top 100 cited results and in future we aim to visualise all retrieved results.

## Acknowledgement

This research has been supported in part by Science Foundation Ireland under Grant Number SFI/12/RC/2289.

## References

1. Chen, H., Sharp, B.M.: Content-rich biological network constructed by mining pubmed abstracts. *BMC bioinformatics* 5(1), 1 (2004)
2. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences LOD Cloud. In: 1st International Workshop on Ontology Engineering in a Data-driven World collocated with EKAW12 (2012)
3. Hasnain, A., Kamdar, M.R., Hasapis, P., Zeginis, D., Warren Jr, C.N., et al.: Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery. In: International Semantic Web Conference (In-Use Track), October 2014 (2014)
4. Hasnain, A., Mehmood, Q., e Zainab, S.S., Decker, S.: A provenance assisted roadmap for life sciences linked open data cloud. In: Knowledge Engineering and Semantic Web, pp. 72–86. Springer (2015)
5. Hasnain, A., e Zainab, S.S., Kamdar, M.R., Mehmood, Q., Warren Jr, C.N., Fatimah, Q.A., Deus, H.F., Mehdi, M., Decker, S.: A roadmap for navigating the life sciences linked open data cloud. In: Semantic Technology, pp. 97–112. Springer (2014)
6. Jonassen, D.H., Beissner, K., Yacci, M.: Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge. Psychology Press (1993)
7. Joseph, T., Saipradeep, V.G., Raghavan, G.S.V., Srinivasan, R., Rao, A., Kotte, S., Sivadasan, N.: Tpx: Biomedical literature search made easy. *Bioinformatics* 8(12), 578 (2012)
8. Kobourov, S.G.: Force-directed drawing algorithms (2004)
9. Lewis, J.R., Sauro, J.: The factor structure of the system usability scale. In: HCD (2009)
10. Monaghan, F., Bordea, G., Samp, K., Buitelaar, P.: Exploring your research: Sprinkling some saffron on semantic web dog food. In: Semantic Web Challenge at the International Semantic Web Conference. vol. 117, pp. 420–435. Citeseer (2010)
11. Osborne, F., Motta, E., Mulholland, P.: Exploring scholarly data with rexplore. In: International semantic web conference. pp. 460–477. Springer (2013)

12. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 990–998. ACM (2008)
13. Wu, Y., Duan, Z.: Visualization analysis of author collaborations in schizophrenia research. *BMC psychiatry* 15(1), 27 (2015)
14. Xuan, W., Dai, M., Mirel, B., Wilson, J., Athey, B., Watson, S.J., Meng, F.: An active visual search interface for medline. In: *Comput Syst Bioinform Conf.* vol. 6, pp. 359–69. World Scientific (2007)
15. e Zainab, S.S., Hasnain, A., Saleem, M., Mehmood, Q., Zehra, D., Decker, S.: Fedviz: A visual interface for sparql queries formulation and execution. In: *Visualizations and User Interfaces for Ontologies and Linked Data (VOILA 2015)*, Bethlehem, Pennsylvania, USA. (2015)
16. Zhao, D., Strotmann, A.: *Analysis and Visualization of Citation Networks*. Morgan & Claypool Publishers (2015)