Big Data and WebGIS for Formulating Health Care Policy in India

Vaibhav Kumar
Data Science and Engineering
Indian Institute of Science Education and Research Bhopal
Madhya Pradesh, India
vaibhav@iiserb.ac.in
https://orcid.org/0000-0002-0047-0681

Ahana Sarkar
Graduate School of Advanced Science and Engineering
Hiroshima University, Japan
ahana@hiroshima-u.ac.jp

Arnab Jana
Centre for Urban Science and Engineering
Indian Institute of Technology Bombay
Mumbai, India
arnab.jana@iitb.ac.in

***Abstract*—Lack of strategic framework in service delivery and decision making remains a void in health care, especially in developing nations. Online social media (OSM) can be an appropriate tool in raising awareness, issuing warnings and retrieving information regarding health issues. This study proposes a web-based health Geographic Information System (GIS) by utilizing a fabrication of OSM sources like government data, news media and Twitter data. Web crawlers were developed to retrieve the historical datasets from the archives of the data sources. The processed data is represented geo-spatially using a developed WebGIS system. The sequential computation-driven system integrating spatial extensions of PostgreSQL database and OpenLayers allows us to analyze health-related information, conduct spatiotemporal queries, and generate spatial density distribution maps to determine the disease hot spots and visualize space-time connections at a local scale. The outcomes can support the agencies in framing healthcare-related policies based on geoinformation intelligence and data analysis on various datasets including social media. It can further pave the way towards an e-governance system for efficient healthcare service delivery to every section of the society.**

Keywords—Health Information System (HIS), Big data, Twitter, WebGIS

I. INTRODUCTION

Recent technological advancements have witnessed the wide penetration of information technology in medicine and healthcare utilizing electronic health records, biomedical database, etc. The inclusion of geoinformation intelligence in health care has been a revelation in understanding the spatial aspects of the events related to health care and epidemics [1]–[3]. However, traditional Geographic Information Systems (GIS) based systems have certain drawbacks such as delays and a lot of effort in information processing and exchange among the stakeholders. This often leads to delayed response and makes it very difficult when real-time information exchange is required. WebGIS-based decision support systems allow online mapping and analysis functionalities, hence can more suitable for health and epidemic-based planning [4]–[6]. It has been further supported by technologies such as cloud computing and big data analytics in developing healthcare cyber-physical systems (CPS) [7].

In the past SQL-based approach has been utilized to control and modify the processing, storage, indexing overlooking the analysis related limitations of healthcare databases [8]. Further, multilingual health search engines have been designed to extract well-structured data from heterogeneous healthcare sources [9]. Apart from developing static health information systems, researchers have also presented dynamic health care systems to store daily information.

Although widespread technologically sound innovation exists in the healthcare field, the holistic social dimension remains a void in healthcare policy formulation, especially in developing nations like India [10]. Apart from data accuracy and false reporting issues, the vital blind spot remains the unawareness concerning the actual ground-level incidences, leading to huge challenges [11]. Even remote stand-alone health-CPS fail to uphold intrinsic functionalities like the socio-demographic dimension, predictive disease capability, disease affected hotspots, and preparedness of forthcoming disease outbreak. In response to this context, big data from Online Social Media (OSM) often aids in creating awareness, sharing information, providing emotional support to the public in different disease and resource delivery systems [12][13]. Another advantage of media-based health information is its spatial dimension and reach capability to urbanites and remote sectors [14]. While the overall socio-demographic situation needs attention, the current dynamic health-CPS lacks spatial tagged information of health incidences. This calls for a web-based dynamic healthcare-CPS design plan that would aid policymakers in decision-making, preparedness, and timely resource allocation.

The recent health-CPS-related literature focuses on aggregating active user input such as smart feedback system, patients' digital health records, biosensors etc., that support the data acquisition for effective decision making. The CPS architecture for healthcare application includes a service-oriented architecture based medical CPS and WSNcloud-integrated secured CPS architecture [15].

This paper aims to improve the health information system in India across all socio-demographic zones such as child health status, vaccination, seasonal variation of disease patterns, etc. Objectively, this study attempts to develop a web-based GIS assisted health informatics portal utilizing OSM data. Also, the advantages and challenges concerning the real-time portal have been revealed here. The lack of literature on the architecture of the OSM-coupled web-based data-driven dynamic portal and its impact on healthcare policy formulation in India is the prior motivation behind this work. Currently limited or no research exists that compares and integrates the data sources, i.e., government portal, twitter and print media, in healthcare-based studies. Hence this is a novelty of the research. The development of the portal would comply with UN-Sustainable Development Goals (SDG) 3, which aims to foster good health for all.

2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW) | 978-1-7281-5384-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/21CW48944.2021.9532567

II. BACKGROUND

A. Indian Health Information Systems: Current status

The dearth of digitized and consistent health data remains a major blind spot in the health platform of India. The National Health Policy (NHP 2017) expressed the need for an improved Health Information System (HIS) with a call for the development of information databases [16]. However, the incomplete and non-digitized census and public health survey data collected and maintained in widely varying database systems compound the problem of HIS formulation. The Health Management Information System (HMIS) portal launched in 2008 aimed to deliver a periodic report on the health status of India. Yet, the HMIS portal currently faces issues such as poor data quality, irrelevant information, duplication of efforts, lack of timeliness and private sector data availability [17]. Furthermore, lack of standardization of data collection and collation in Indian hospitals leads to inter-district and inter-state level data quality differences, thereby resulting in a dearth of data sharing, duplication of efforts, data inaccuracies, delays in detection, delays in response time and variety in quality state health delivery systems [18]. Dynamic problem detection methods are underemphasized, and to the extent they exist, but are not integrated with other systems.

B. Use of OSM and print media for HIS

Recent advancement in health-based technology has triggered the usage of social media platforms. Historical and real-time data can be collected and analysed from social media like Twitter to extract meaningful information [12]. While social media has been used in health sciences for creating awareness, sharing information and providing emotional support to the public in different diseases, on the one hand, it has been employed as a tool for resource delivery in post-disaster medical treatments [13]. Furthermore, media-based health information reaches not only the urbanites but also the population in remote areas. Among many advantages, the major impacts of OSM data on health include better access to documented health records, improved recommendations, personalized health services, emergency medical response, etc. [19]. Hence, there is an urgent need for modern and improved application of WebGIS based online system.

This study utilizes different online data sources like government websites, print media and Twitter data to develop a WebGIS health informatics portal.

III. METHODOLOGY

The methodology initiated with the development of a web-based GIS system for big-data analytics in health care decision making utilizing a sequential computation-driven approach. Fig.1 showcases the development process and the components of the WebGIS system.

A. Twitter, print media and government data preparation

In this study, three data sources have been considered for demonstrating the usefulness as well as challenges of utilising big data in healthcare which are as follows:

a) Print Media: Time of India archives
b) Twitter
c) Government crowdsourced data sources

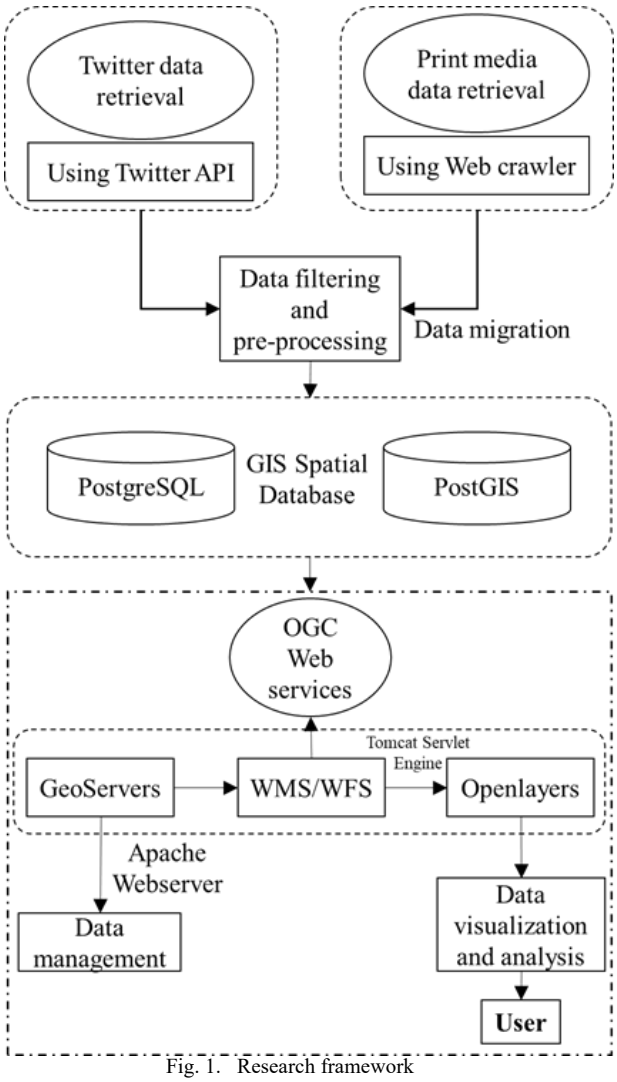

Fig. 1. Research framework

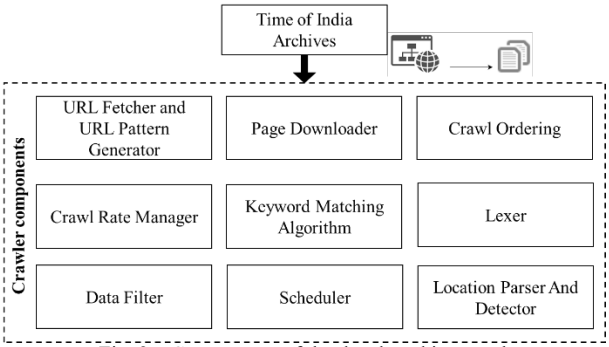

Fig. 2. Components of the developed internet bot

Print media web crawler or internet bot was utilized to systematically browse the World Wide Web, and extract specific HTML data from websites based on the desired keywords in the URLs [12]. Fig.2 illustrates the components of the developed web crawler. Crawling speed varies upon the system configuration, bandwidth and noise in the data. One of the noise sources is the unwanted tags or text that get scraped along with the relevant data. Cleaning up the data turn to be a challenging task as the website gets frequently updated; thus, scrapping needs to be updated very frequently. A python module, *Scrapy* was used to implement the crawler for online archives of the newspaper "The Times of India". Keywords/seeds based search was performed to enlist the

initial URLs. The hyperlinks from the visited URLs were added to the URL queue to form the crawler frontier followed by the recursive visit of the selected URLs and parsing of the data. The tweets and the print media text were parsed to filter and retrieve the time and location information using text parsing. The location information was retrieved using the Google Geocoding API service by parsing the parsed text. The data was then filtered based on the acquired location information. We have used APIs provided by Twitter to extract the historical tweets.

TABLE 1. ACQUIRED DATA TABLES

Source	Data count							
	2018				2019			
	Q1	*Q2*	*Q3*	*Q4*	*Q1*	*Q2*	*Q3*	*Q4*
Print media	16	26	13	18	13	15	13	38
Twitter	12	18	11	14	13	20	16	18
Gov. data	14	15	17	16	16	12	13	23

Based on the trending topics collected from the website "trend calendar[1]" following keywords were used to retrieve the data from both the sources: *Dengue, Vector-borne disease, Kala-azar, Malaria, immunization, Vaccination, Mother and child vaccination, Health, Healthcare, Healthcare India, Vaccination program, Rural vaccination, Rural healthcare.*

B. Data migration to spatial database

The filtered data was then migrated to a spatial database due to the available functionalities, efficient large and varying spatial data handling, and redundancy-reduction capacity. PostgreSQL database and Post-GIS as its spatial extension were used for data management. This DBMS is freely available for the users and offers a lot of spatial functionalities to handle and perform analysis on the datasets.

C. Data visualization and analysis on WebGIS portal

The Web GIS is an extension and application of traditional client/server computing, in which the geospatial data is accessible in a shareable environment. The web-enabled GIS assist decision making at the strategic and operational levels. It further supports administrative operations for decision-makers and general users to access the information conveniently and effectively. Also, WebGIS makes it easy to access and acquire GIS data from diverse data sources in the distributed environment. In this research, a Web-based GIS system was developed for visualization and analysis of the spatial data. The system will perform analysis such as on the fly user defined distance-based heat map development. Apart from developing heat maps on historical data, one unique feature of the system can dynamically change the heat maps based on updated data extracted from sources such as tweets.

Geoserver, a Java-script based open-source webserver, was selected for its capability to share, process, control and publish the geo-spatial data as Web Map Service (WMS) and Web Feature Service (WFS). OpenLayers, an open-source JavaScript library was used to display and manipulate map data in a GIS functionality-enabled client-side web browser. It provides an application programming interface (API) for building rich web-based geographic applications similar to Google Maps and Bing Maps [21], [22]. Hence the CPS architecture of the web-based system was developed by integrating OpenLayers, PostGIS, and Geoserver. This integrated CPS was capable of analyzing health-related information, conducting temporal and spatial queries, generating spatial density distribution maps across a region to determine the occurrence of hot spots, and visualizing space-time connections at a local and regional scale.

The data was analyzed in various spatiotemporal scales. The temporal scale was accounted for i) two years and ii) four quarters per year. Table I details the number of geocoded data points for the sources.

IV. RESULTS AND DISCUSSION

It can be observed that there are significant variations in the number of retrieved data from each source across the periods. Nevertheless, all these datasets lacked granular-level information. The spatial pattern of the datasets was investigated further to gain insight into the variations in the hotspots. Fig.3 elucidates the WebGIS interface, which comprises the various functionalities utilized to filter and visualize the data. Furthermore, different base-maps of "Statmen" can be selected to thematically represent the location clusters as location points and heat maps, respectively. The data information window has been provided to view the data information. The dynamic feature of real-time Twitter live feed window has been added to show the latest updated tweet based on the provided tweet hashtags (see Fig.3).

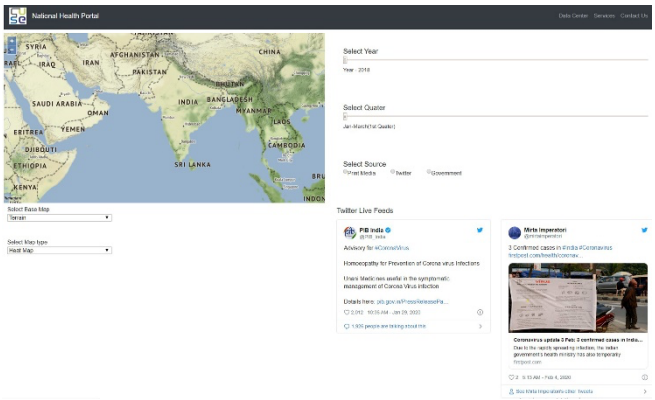

Fig. 3. WebGIS interface with terrain base map (Map source: Stamen[2])

Fig. 4 elucidates the heat maps, a thematic representation of location clusters for the print media, Twitter, and government data acquired for the four quarters of the years 2018 and 2019. While Fig. 5 represents the spatiotemporal variations in the print media across the quarters for the year 2019.

A clear variation in the hot spots confirms the non-uniformity in the datasets from the recorded sources. The hotspot variations were also observed in every quarter for both the year. Most of the data locations of print media and twitter did not match the government data sources. This is mostly due to the limitation in the reporting. For example, a majority of print media data was reported at the district or state scale.

Fig. 4 also elucidates the maximum presence of Twitter data. Nevertheless, the tweets mostly comprised of the reactions to the print media. This can be attributed to various reasons; one of them could be the dearth of uniformity in reporting issues,

[1] https://us.trend-calendar.com/trend/2020-08-01.html

[2] http://maps.stamen.com

resources and limited ICT usage across remote sectors of India. Another significant aspect is eradicating regional languages in data while retrieving information advertently leading to data losses. Addressing this limitation remains future work. The data location reported by every media is different, a challenge arises regarding the authority of the various datasets. The results indicate that the data reporting and data repository require a predefined framework to maintain data consistency and reduce data redundancy, which requires further investigations. Furthermore, the accuracy levels of the determined locations need to be intensified for better decision making. This challenge can be addressed by digitation of the reported values and the location information by the government agencies. The data can be used to validate the datasets of social media sources such as Twitter. Another solution could be using predefined Twitter handles and hashtags by assigned a human resource at every rural, zonal, regional, and national health centers. These resources would report the credible information such as details of patients, resource availability, and health cases using Twitter. Such datasets can be analyzed in real-time using the proposed systems combined with Natural Language Processing based algorithms. By combining it with the proposed WebGIS-based online portal quick decision-making can be done apart from formulating policies. It could also help citizens to be aware of unprecedented situations. Although the number of Twitter data was less than print media sources, it can prove to be a vital source of health data.

Integrating data with geostatistical analysis would provide eloquent spatiotemporal patterns and relationships. If dispensed with quality digital data, the geoportal can significantly enhance the utilization of reckoned social media data for effective policy formulation.

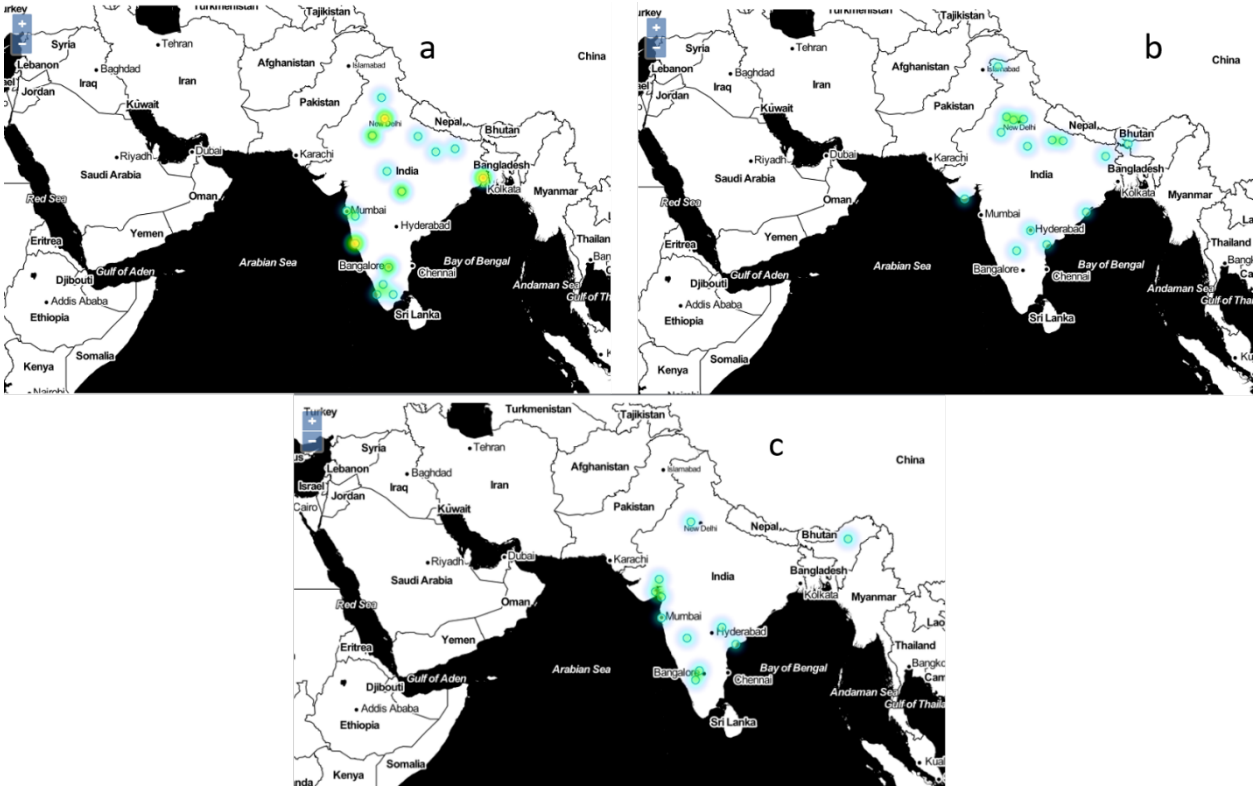

Fig. 4. Variations in Heatmap for (2019, January to April) a) print media data b) twitter c) government data (Map Source: Stamen[3])

[3] http://maps.stamen.com

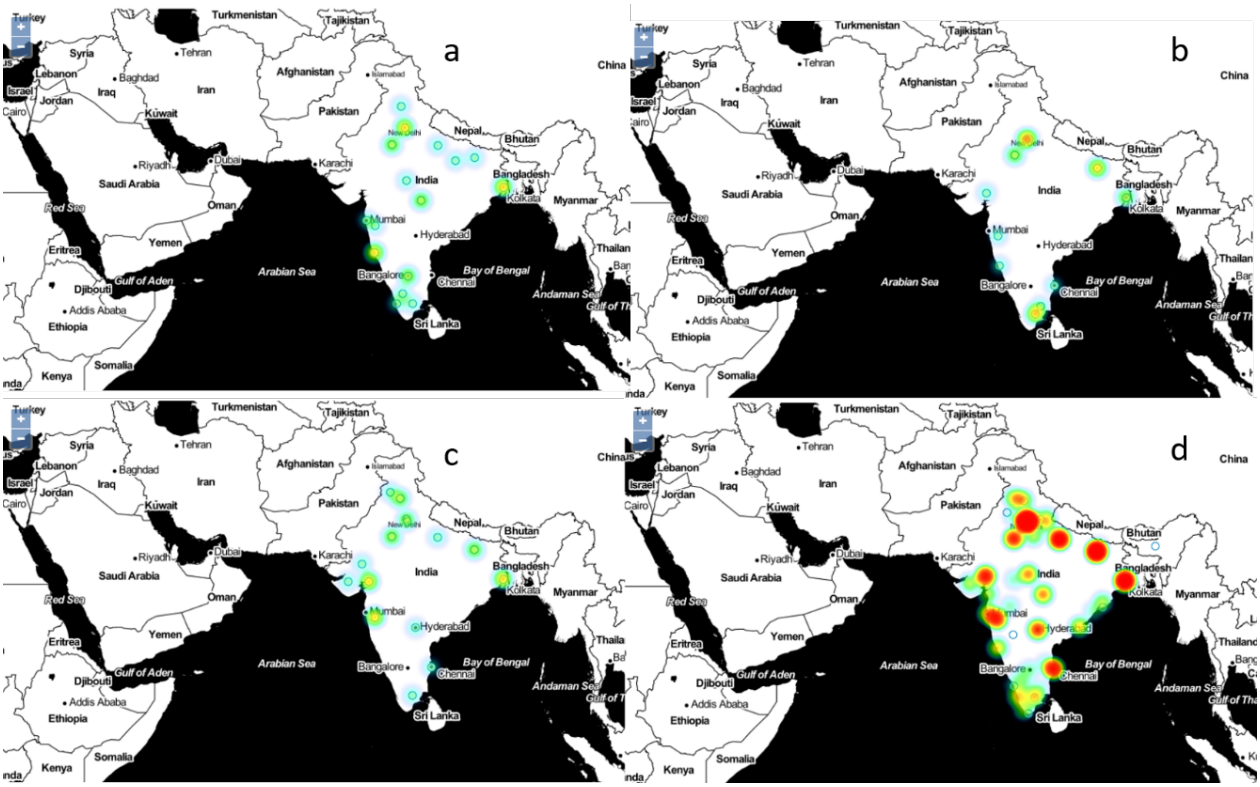

Fig. 5. Variations in Heatmap for in print media data for four quarters of 2019 (Map Source: Stamen[4])

V. Conclusion

Health infrastructure delivery poses a major challenge for government authorities. In comparison to the current trend of application of social and print media in answering various issues like disaster and sentiment analysis, an obvious blind spot was noted in its application in the health sector. The challenge to fill the gap between supply and demand of health facilities is primarily due to the lack of credible data for policy formulations. Huge locational variations in the datasets acquired from different data sources is another challenge. This raise concerns over the collective credibility of the datasets. To address the concerns, the proposed system can be utilized to acquire credible information through appointing dedicated resources that could share real-time information regarding health issues using predefined Twitter handle and standard information formats. This will also establish a credible synchronization between the considered data sources at the regional, district, and rural level.

This exercise would not only develop a real-time structured database but would also intensify resource-delivery system. Integrated big data-GIS approach would further aid in identifying the event trends and spatiotemporal patterns through spatial map generation. The outcomes of this study turn out to be an integrally crucial component of the e-governance system in addressing healthcare delivery-related issues. The data in the social domain is of wide variety and veracity. We have extracted the data using keywords without language processing, which might have resulted in data loss. Hence future work involves the development of learning and language processing classification methods. Moreover, Artificial Intelligence (AI) based algorithms can be developed to analyze the sentiments of the citizens for the health related policies and activities [23].

Acknowledgment

Part of this work is funded by the Interdisciplinary Cyber Physical Systems (ICPS) programme by Department of Science and Technology (DST), Government of India; Project Grant number: RD/0118-DST0001-002. Any options, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DST.

References

[1] E. K. Cromley, “Using GIS to Address Epidemiologic Research Questions,” Curr Epidemiol Rep, vol. 6, no. 2, pp. 162–173, Jun. 2019, doi: 10.1007/s40471-019-00193-6.

[2] B. F. Khashoggi and A. Murad, “Issues of Healthcare Planning and GIS: A Review,” ISPRS International Journal of Geo-Information, vol. 9, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/ijgi9060352.

[3] R. Roquette, M. Painho, and B. Nunes, “Spatial epidemiology of cancer: a review of data sources, methods and risk factors,” Geospatial Health, vol. 12, no. 1, p. 504, May 2017, doi: https://doi.org/10.4081/gh.2017.504.

[4] S. Verma, “Development of Web GIS Based Framework for Public Health Management System Using ERDAS Apollo 2010,” 2017.

[5] A. R. A. Rasam, A. H. Azlin, and N. M. Saraf, “Mobile Apps and Web GIS-Based Accessible Health and Social Care System for People with Disabilities,” in 2018 IEEE 8th International Conference on System Engineering and Technology (ICSET), Oct. 2018, pp. 85–90, doi: 10.1109/ICSEngT.2018.8606358.

[6] S. Mansour, “Spatial analysis of public health facilities in Riyadh Governorate, Saudi Arabia: a GIS-based study to assess geographic variations of service provision and accessibility,” Geo-spatial Information Science, vol. 19, no. 1, pp. 26–38, Jan. 2016, doi: 10.1080/10095020.2016.1151205.

[4] http://maps.stamen.com

[7] Y. Zhang, M. Qiu, S. Member, and C. Tsai, "Health-CPS : Healthcare Cyber-Physical System Assisted by Cloud and Big Data," IEEE Systems Journal, vol. 11, no. 1, pp. 88–95, 2017, doi: 10.1109/JSYST.2015.2460747.

[8] C. Lin, L. Huang, and S. T. Chou, "Temporal Event Tracing on Big Healthcare Data Analytics," 2014 IEEE International Congress on Big Data, no. M, pp. 281–287, 2014, doi: 10.1109/BigData.Congress.2014.48.

[9] L. Nie, T. Li, M. Akbari, J. SHEN, and T.-S. CHUA, "WenZher : Comprehensive Vertical search for Healthcare Domain," in Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information, 2014, pp. 1245–1246.

[10] R. Gaitonde, M. San Sebastian, V. R. Muraleedharan, and A.-K. Hurtig, "Community Action for Health in India's National Rural Health Mission: One policy, many paths," Social Science & Medicine, vol. 188, pp. 82–90, Sep. 2017, doi: 10.1016/j.socscimed.2017.06.043.

[11] O. S. Albahri et al., "Systematic Review of Real-time Remote Health Monitoring System in Triage and Priority-Based Sensor Technology: Taxonomy, Open Challenges, Motivation and Recommendations," J Med Syst, vol. 42, no. 5, p. 80, Mar. 2018, doi: 10.1007/s10916-018-0943-4.

[12] Z. Zhang and A. Wasim, "A comparison of information sharing behaviours across 379 health conditions on Twitter," International Journal of Public Health, vol. 64, no. 3, pp. 431–440, 2019, doi: 10.1007/s00038-018-1192-5.

[13] R. Basu, A. Khatua, A. Jana, and S. Ghosh, "Harnessing Twitter Data for Analyzing Public Reactions to Transportation Policies : Evidences from the Odd-Even Policy in Delhi , India," no. April 2018, 2017.

[14] S. Jahan, N. Hasan, R. Hasan, and J. Gammack, "Healthcare support for underserved communities using a mobile social media platform," Information Systems, vol. 66, pp. 1–12, 2017, doi: 10.1016/j.is.2017.01.001.

[15] S. A. Haque, S. M. Aziz, and M. Rahman, "Review of Cyber-Physical System in Healthcare Review of Cyber-Physical System in Healthcare," International Journal of Distributed Sensor Networks, no. May, 2014, doi: 10.1155/2014/217415.

[16] K. P. Principles et al., "NATIONAL HEALTH POLICY , 2017."

[17] J. Samal and R. K. Dehury, "Perspectives and Challenges of Hmis Officials in the Implementation of Health Management Information System (HMIS) with Reference to Maternal Health Services in Assam," Journal of Clinical and Diagnostic Research, vol. 10, no. 6, pp. 9–11, 2016, doi: 10.7860/JCDR/2016/16921.7955.

[18] K. Valley, A. Nori-sarma, A. Gurung, G. S. Azhar, and A. Rajiva, "Opportunities and Challenges in Public Health Data Collection in Southern Asia : Examples from," Sustainability, vol. 9, 2017, doi: 10.3390/su9071106.

[19] M. Thomas and P. Narayan, "Information Technology for Development The Role of Participatory Communication in Tracking Unreported Reproductive Tract Issues in Marginalized Communities," vol. 1102, 2016, doi: 10.1080/02681102.2014.886549.

[20] Y. Kim and Y. Kim, "Implementation of hybrid P2P networking distributed web crawler using AWS for smart work news big data," Peer-to-peer Networking and Applications, 2019.

[21] H. C. Karnatak et al., "Spatial mashup technology and real time data integration in geo-web application using open source GIS – a case study for disaster management," Geocarto International, vol. 27, no. 6, pp. 499–514, 2012, doi: 10.1080/10106049.2011.650651.

[22] E. M. Delmelle, H. Zhu, W. Tang, and I. Casas, "A web-based geospatial toolkit for the monitoring of dengue fever," Applied Geography, vol. 52, pp. 144–152, 2014, doi: 10.1016/j.apgeog.2014.05.007.

[23] A. Giahanou and F. Crestani, "Like It or Not," ACM Comput. Surv. vol. 49, no. 2, pp. 1-41, 2016, doi: 10.1145/2938640.