Check for updates

RESEARCH ARTICLE

REVISED A data-driven global observatory addressing worldwide challenges through text mining and complex data visualisation

[version 2; peer review: 1 approved, 3 approved with reservations]

Joao Costa[1,2], M. Besher Massri[2,3], Marko Grobelnik[1-3], Ignacio Casals del Busto[4], Dale Weston[5]

[1]Quintelligence, Ljubljana, Slovenia
[2]UNESCO International Research Institute on AI - IRCAI, Ljubljana, Slovenia
[3]Institute Jozef Stefan, Ljubljana, Slovenia
[4]Aguas del Alicante, Alicante, Spain
[5]Public Health England, Salisbury, UK

v2 **First published:** 30 May 2022, **2**:68
https://doi.org/10.12688/openreseurope.14471.1
Latest published: 07 Dec 2023, **2**:68
https://doi.org/10.12688/openreseurope.14471.2

Abstract

Observing the world on a global scale can help us understand better the context of problems that engage us all. In this paper, we propose a data-driven global observatory methodology that puts together the different perspectives of media, science, statistics and sensing over heterogeneous data sources and text mining algorithms. We also discuss the implementation of this global observatory in the context of epidemic intelligence, monitoring the impact of the COVID-19 pandemic, and in the context of climate change, with a specific focus on water resource management. Moreover, we discuss the value of this global solution in local contexts and priorities, based on the exchange with stakeholders in municipalities, utilities and governmental institutions.

Plain language summary

In this research paper we lay the foundations for a data-driven perspective on how to appropriately use open data to build business intelligence, with proven usefulness in applications to public health and water resource management and with potential applications to other priority areas.

Keywords

Big Data, Semantic Technologies, Public Health, Water Management, Text Mining, MeSH Headings, MEDLINE, multilingual news, Social media

Open Peer Review

Approval Status ? ✓ ? ?

	1	2	3	4
version 2 (revision) 07 Dec 2023		✓ view	? view	? view
version 1 30 May 2022	? view	? view		

1. **Haowen Xu**, Oak Ridge National Laboratory, Bethel Valley Road, USA
2. **Carson Leung**, University of Manitoba, Manitoba, Canada
3. **Farah Jemili**, Universite de Sousse, Sousse, Tunisia
4. **Ángel M. García-Vico**, University of Jaén, Jaén, Spain

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Horizon 2020 gateway.



This article is included in the Data Science gateway.



This article is included in the Research on Research gateway.

Corresponding author: Joao Costa (joao.pitacosta@quintelligence.com)

Author roles: Costa J: Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Massri MB**: Investigation, Resources, Software, Visualization, Writing – Review & Editing; **Grobelnik M**: Conceptualization, Investigation, Methodology, Supervision, Writing – Review & Editing; **Casals del Busto I**: Validation, Writing – Review & Editing; **Weston D**: Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreements No 727721, A holistic water ecosystem for digitisation of urban water sector (NAIADES) and No 820985, Meaningful Integration of Data, Analytics and Services (MIDAS).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2023 Costa J *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Costa J, Massri MB, Grobelnik M *et al.* **A data-driven global observatory addressing worldwide challenges through text mining and complex data visualisation [version 2; peer review: 1 approved, 3 approved with reservations]** Open Research Europe 2023, **2**:68 https://doi.org/10.12688/openreseurope.14471.2

First published: 30 May 2022, **2**:68 https://doi.org/10.12688/openreseurope.14471.1

REVISED **Amendments from Version 1**

After a useful peer review, we have improved the content taking into consideration the comments that mostly regarded the context to related work. We have improved the readability of this paper adding new references to this work, and updated the context in which it was presented when first submitted.

Any further responses from the reviewers can be found at the end of the article

II. Introduction

The world's globalization phenomena unveiled awareness of worldwide problems, such as the climate crisis, but also to common efforts to find solutions to those problems, as was the case of the several COVID-19 collaborative actions. There are many obstacles still today on such global strategies to which innovative technology and data-driven solutions can help. In this paper we propose the concept of a global observatory based on text mining algorithms that is able to answer the wide range questions that are core to global solutions, using machine learning and Big Data analytic methods over the layered information it is ingesting often in real-time. The main perspectives of this global observatory fall on: (i) the monitoring and exploration of news articles and social media feeds; (ii) the analysis of combinations of indicators through time and what stories can they tell; and (iii) the exploration of published scientific knowledge. All of these perspectives can be combined to provide complementary answers to main topics from health to engineering. In this paper we discuss the results obtained based on two implementations of these observatory approach: (a) the Coronavirus Watch portal released in 1 addressing the worldwide spread of COVID-19[2]; and (b) the NAIADES Water Observatory[3], focusing on best practices to build water sustainability.

III. Methods

Taking into account the schema in Figure 1, we consider the construction of the Global Water Observatory into phases, going from lower to higher complexity. A similar observatory, dedicated to monitor COVID-19[2], was made available with less functionality but also including a diversity of perspectives for which the interoperability is a core topic of discussion in this paper. We start by putting together data sources that are meaningful to a range of stakeholders targeted within engaged citizens to decision makers and that can leverage the information provided to established evidence-based policies.

At the data collection phase, we are concerned with addressing properly the challenges in the heterogeneous nature of the data, their different frequency and size, as well as the levels of access to it established by data providers. These parameters to take into consideration ensure the appropriate data ingestion into the system. The selection of data sources and features to be ingested is done manually, but the ingestion itself is automated when the frequency requires so. The frequency of update depends solely of the data provider. At this first stage we are collecting data from many different data sources (such as, *e.g.*, the Worldwide news, the Microsoft Academic Graph, the Word Bank, the United Nations Sustainable Development Goals), according to their relation to the focus topics and priorities.

A forthcoming stage is in the data cleaning, data processing and data integration prior ingestion. This step is highly important to allow for the data quality that is needed in order to obtain useful insights from it. In this step we include the data curation, where the most meaningful datasets are selected and parsed. We also include the exploratory data analysis and some data visualisation for the purpose of prototyping what is then available at the Water Observatory.

The Observatory phase is then possible when the curated data streams of a selection of dynamic data sources are live in the system and can be used to obtain insight on particular topics of interest, monitor Key Performance Indicators associated with business priorities, and allow for a global and local perspective on related topics. These include interactive data visualisations of indicators and statistical data, the dynamic view of the news sources over priorities, or the user query over a scientific research topic. This allows for insight on topics in analysis (such as water topics like, *e.g.*, water scarcity and water quality, and public health topics like, *e.g.*, ebola or the new coronavirus) that will be put into the context of local data when sourced from the shared interest of users.

The path ahead is a novel concept of a meaningful Digital Twin (*i.e.*, a dynamical model which, given a current state of an observed system, is capable of a digital partial reconstruction of such a system) that builds over the Global Water Observatory to rise above data complexity towards data interoperability. This is usually difficult to achieve in full due to the heterogeneity of the data, the different characteristics of the data sourced (frequency, data types, *etc.*) and the domain knowledge needed to identify new challenges covering a wide range of

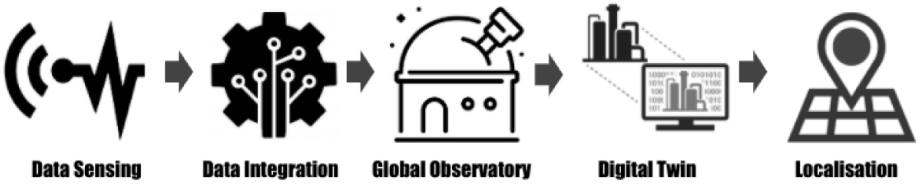

Figure 1. The approach used leading from data sensing to the digital twin and its approximation to local priorities.

business intelligence priorities. Nevertheless, useful aspects of it can be achieved, some of which are already evident from the implementations discussed in this paper. An example of this is to track a topic in the news, its impact in the social media, and explore the range of the problem in the published scientific research, as well as extract good practices to deal with this problem.

We add a final stage to this diagram that is usually forgotten in a theoretical framework, which is the adaptation of the system to the needs and priorities in the user side. Here we consider the ingestion of local data, the customization of news streams, the availability of exploratory dashboards, the shareable instances for policy makers, and the APIs for 3rd party integration.

The system that is able to access the data sources that relate to the items above, is also able to track the term throughout the several phases of popularization. It is also able to show the current status of a particular topic of interest, and optimally alert for potentially trendy topics in the future. In that particular context, the interactive data visualisation is a key factor to improve the usefulness of the tool and should express visual narratives that comprehend the relevant aspects of the problem. Good examples of such can be consulted for the case of epidemic intelligence in 4 and water intelligence in 5, and will be discussed in the implementations described and explored for the purpose of this study.

IV. On The Coronaviruswatch Observatory

When the World Health Organization (WHO) announced the global COVID-19 pandemic on March 11th 2020[6], following the rising incidence of the SARS-CoV-2 in Europe, the world started reading and talking about the new Coronavirus. The arrival of the epidemic to Europe scaled out the news published about the topic, while public health institutions and governmental agencies had to look for existing reliable solutions that could help them plan their actions and the consequences of these.

Technological companies and scientific communities invested efforts in making available tools (*e.g.* the GIS[7] later adopted by WHO), challenges (*e.g.* the Kaggle COVID-19 competition[8]), and scientific reports and data (*e.g.* the repositories medRxiv[9] and Zenodo[10]).

In March 2020 we have released the first implementation of this global observatory as the Coronavirus Watch portal[1], aiming to contribute to a multinational response to the global crisis. It was made available by the UNESCO AI Research Institute (IRCAI), comprehending several data exploration dashboards related to the SARS-CoV-2 worldwide pandemic. This platform aimed to expose the different perspectives on the data generated and trigger actions that can contribute to a better understanding of the behavior of the disease (see Figure 2).

The portal includes a real-time news monitoring system that can be focused at European and national level, side-by-side with the data on the progress of the pandemics made available by the Worldometer[11] and the Center for Disease Control[12]. The visual representation of the details of those indicators were made available over animations showcasing the live comparisons in 5D (as in Figure 2), the trajectories of the most affected

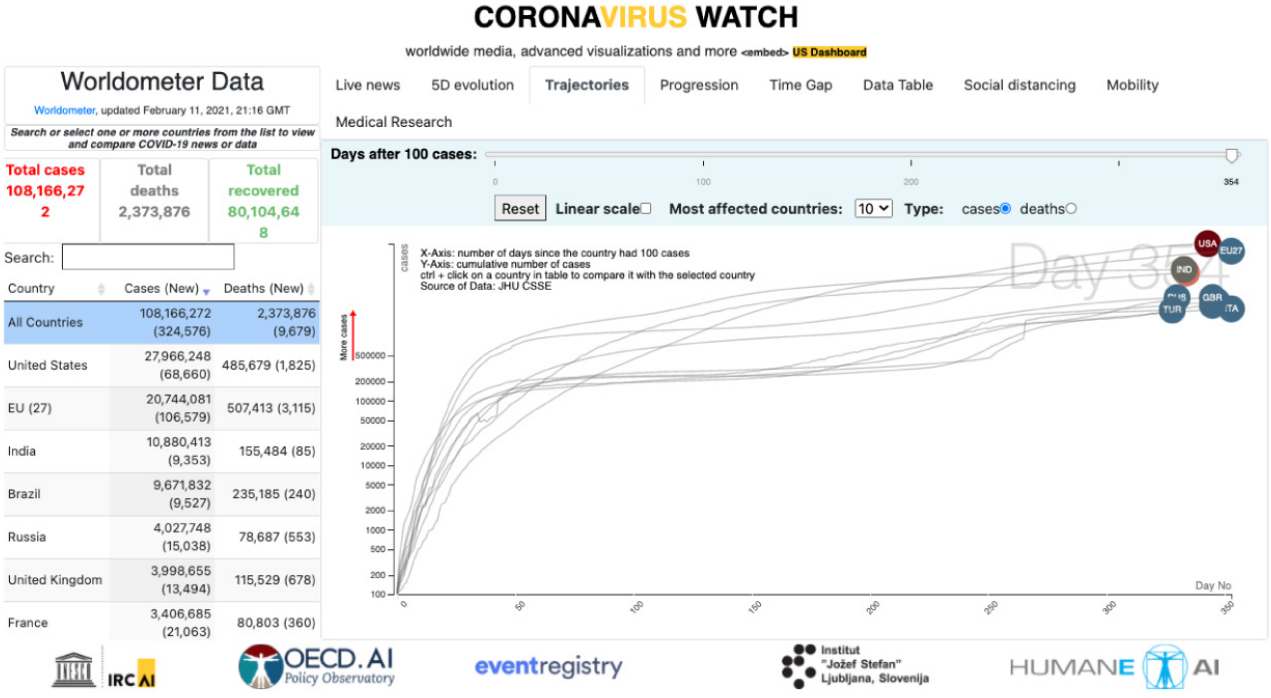

Figure 2. The IRCAI observatory for the COVID-19 pandemic.

countries, and the details of the progression of the disease. It also included perspectives on the mobility, sourced on the Google Community mobility data[13], a social distancing simulator, and exploration tools based on the published biomedical research (see 2).

To improve the resolution of these results and to optimise their relevance for European public health agencies, we have developed a set of COVID-19 focused tools on the MIDAS platform[14]. This system was designed for evidence-based decision-making in public health. This approach allows us to validate the usability of the global observatory on a cross-EU level within the COVID-19 context, integrating both health news and biomedical research exploration (see Figure 3).

V. On The Water Observatory

Climate change is a global problem that in the recent years has been in the focus of European and Worldwide strategies. The priorities in European Union are rapidly changing towards sustainability and environmental efficiency, transversely to most domains of action. The European Commission's Green Deal aiming for a climate neutral Europe in 2050, and boosting economy through green technology[15] provides a new framework to understand and position water resource management in the context of the challenges of tomorrow[16]. In this context, the NAIADES project[17] aims to improve the water resource management in a global context, including European regions where water scarcity is predicted, also dealing with concerns as, *e.g.*, saline intrusion and groundwater contamination. To contribute to this cause, we deployed a Global Water Observatory[3] that is focusing on water-related aspects allowing the user to explore the several layers of information it is providing, from news and social media to published science, weather models and indicators. The NAIADES Global Water Observatory does not only contribute to the improvement of European sustainability in water-related matters, but also provides the local actors on the water resource management an active role in that taking into consideration the national and international open data available on water resource management-related topics and priorities (see 18 for more details).

Water is fundamental to all human activity and ecosystem health, and is a topic of rising awareness in the context of the recent discussions on climate change. Water resource management is central to those concerns, with the industry accounting for over 19% of global water withdrawal, and agricultural supply chains are responsible for 70% of water stress[19]. In 2015 the UN established "clean water and sanitation for all" as one of the 17 Sustainable Development Goals, aiming for eight targets to be achieved by 2030[20]. The UN secretary-general points out in April 2020 that SDG 6 is "badly off track" compromising the progress on the 2030 Agenda[21]. As noted by the Organisation for Economic Cooperation and Development (OECD), the 'water crisis' has often proven to be a crisis of governance[22], where water scarcity is largely caused by mismanagement of resources, leading to a global prioritisation[23].

The intention to globally monitor water resources is not new, and already in the late 1960s[24] the first spatially-distributed water resources model appeared, with first operational uses of satellite observations in water resources developed in the early 1980s[25]. The reliable management of water resources is only possible under condition of availability of adequate qualitative and quantitative information about state of the water body at any moment of time. Taking advantage of the recent technological progress enabling much innovation that was

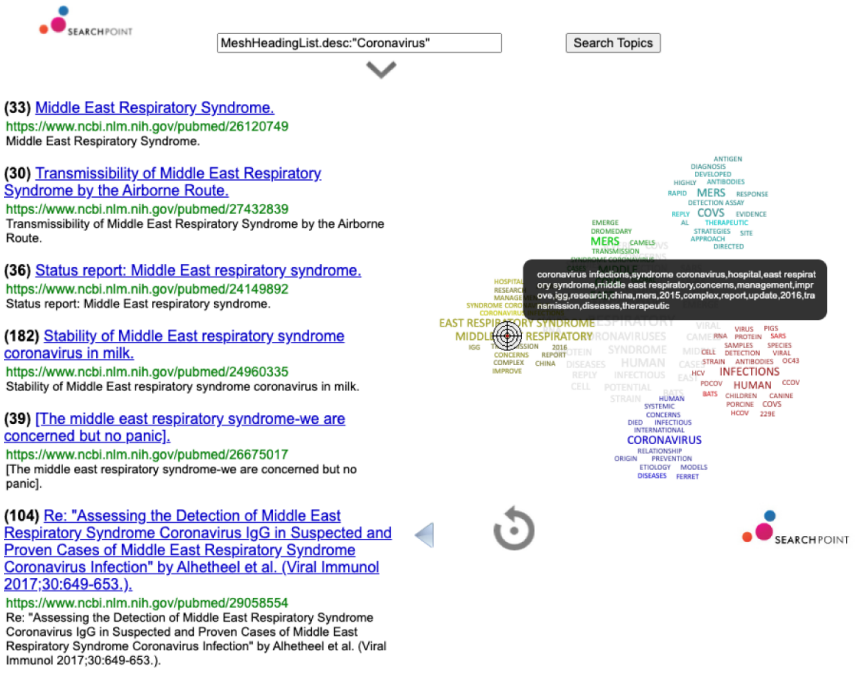

Figure 3. Exploring scientific research through complex data visualisation.

unthinkable a few years ago, the concept of the Digital Twin is increasingly entering the water sector as an innovation driver. Due to the rapidly growing awareness of the sustainability challenges that we are facing in Europe and worldwide in the context of the water resource management, there has been much work done to develop systems that are able to collect information about the available water and even simulate and forecast that in the near future. These are usually geolocation-based systems ingesting water-related data to enable real-time monitoring of resources and usage[26]. The other typical approach is the systems focused on workflows in the water sector, including the management of water distribution networks, hydraulic efficiency or leak/fraud detection, better suited to those companies that already have their infrastructure in place and know well what do they want to monitor[27].

The approach we proposed in this paper is novel in many ways. The news monitoring perspective is monitoring water scarcity and water quality in worldwide and, in particular, in the surrounding regions of the water resource management agencies is is mainly addressing, together with their audiences. This is also including a Twitter observatory that adds to the already implemented measure of impact of the monitored news in Facebook, a social media component to the observatory. The global indicators that are already available for visualisation, sourced over the UN Open Data Portal, the water-focused Sustainable Development Goal 6, and the World Bank Data Portal, can help us understand water-related aspects of the climate crisis.

The important role of scientific research in this context, and the best practices that can be extracted from this data, is explored with a complex data visualisation technology that allows the user to powerful Lucene-based queries over the article's metadata aiming to refine search by moving a pointer over clusters of related topics (see Figure 3). We will also be including other data analytics technologies to analyse simultaneously multiple time-series providing interactive exploration tools to understand trends in the weather and water-related impacts to it.

The localisation of this global system entails the customisation of its functionality in news monitoring, ingestion of local indicators and exploration of scientific research on observed problems in, *e.g.*, groundwater contamination. In that, the observatory is synchronising with the priorities of regional water providers. These agencies (*e.g.* Aguas de Alicante) are collecting data on their water resource management services to improve the customer satisfaction and optimise their system.

VI. Conclusions and future work

The results discussed in this paper show the potential impact of the proposed data-driven global observatory in contexts like public health and climate change preparedness. This integrated system is capable to monitor in real-time the worldwide news and social media, statistics, published science, weather and many other data streams that are identified as useful and can be provide complementary value to those considered already. We will be deploying this system in the context of other global problems where there is enough data to provide useful and meaningful contribution, either in other aspects of the climate crisis to better plan response, in addressing other epidemiological concerns to serve as early warning, or in addressing a new focus in the context of data science for social good.

We are now working on extending this system to integrate the information retrieved by the topics searched over the internet provided by Google Trends, regarding issues related to the context in focus. The user will be able to explore a wide range of indicators and compare trends in a global and local level throughout a meaningful timeline. We will also be reusing EC-funded open datasets and initiatives in order to ingest this information as European-level indicators to complement the analysis. Furthermore, we will be further investigating the validity of the localization of this Global Water Observatory, integrating some of the local data that can be provided by the user, and customizing news sources to their own priorities, as well as making available data exploration dashboards that allow for further insight and evidence-based policy.

VIII. Ethics and consent

Ethical approval and consent were not required

VII. Data availability

For this paper, we used only open data. In particular, we have used the MEDLINE dataset[28] and the worldwide news are being collected by 29, freely available online, but which the dataset we do not have permission to share.

References

1. IRCAI: **Coronavirus Watch portal**. 2020.
Reference Source

2. Massri B, Costa JP, *et al.*: **A global COVID-19 observatory, monitoring the pandemics throughtext mining and visualization.** *Informatica*, Slovenian Society Informatika, 2021; **46**(1).

3. **NAIADES Water Observatory**. 2020.
Reference Source

4. Leung CK, Chen Y, Hoi CSH, *et al.*: **Big data visualization and visual analytics of COVID-19 data.** In: *2020 24th International Conference Information Visualisation (IV).* IEEE, 2020; 415–420.
Publisher Full Text

5. Xu H, Berres A, Liu Y, *et al.*: **An overview of visualization and visual analytics applications in water resources management.** *Environ Model Softw.* 2022; **153**: 105396.
Publisher Full Text

6. World Health Organization: **WHO Director-General's opening remarks at the media briefing on COVID-19 -11 March 2020**.
Reference Source

7. ArcGIS: **WHO COVID-19 Dashboard**. 2020.
Reference Source

8. Kaggle: **COVID-19 Open Research Dataset Challenge**.
Reference Source

9. medRxiv: **COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv**. 2020.
Reference Source

10. **Zenodo Coronavirus Disease Research Community**. 2020.
Reference Source

11. **WorldoMeters**. 2020.
Reference Source

12. **Center for Disease Control and Prevention**. 2020.
Reference Source

13. Google: **COVID-19 Community Mobility Report**. 2020.
Reference Source

14. Pita Costa J, Grobelnik M, Fuart F, *et al.*: **Meaningful Big Data Integration For a Global COVID-19 Strategy.** *Computer Intelligence Magazine.* IEEE, 2020; **15**(4): 51–61.
Publisher Full Text

15. European Commission: **European Green Deal**. 2019.
Reference Source

16. European Commission: **Water Scarcity and Droughts in the European Union**. 2019.
Reference Source

17. Costa JP, Massri MB, Novalija I, *et al.*: **Observing Water-Related Events for Evidence-Based Decision-Making.** In: *The Porceedings of the 2021 Slovenian KDD Conference.* Institute Jozef Stefan.
Reference Source

18. Varady RG, Albrecht TR, Gerlak AK, *et al.*: **Global water initiatives redux: A fresh look at the world of water.** *Water.* 2022; **14**(19): 3093.
Publisher Full Text

19. Our World in Data: **Water Use Stress**. 2020.
Reference Source

20. United Nations Development Programme: **Goal6 -clean water and sanitation**. 2020.
Reference Source

21. Blazhevska V: **United Nations launches framework to speed up progress on water and sanitation goal.** United Nations Sustainable Development, 2020.
Reference Source

22. OECD: **Water governance in OECD countries: A multi-level approach**. OECD, 2011.
Reference Source

23. Akhmouch A, Clavreul D, Glas P: **Introducing the OECD principles on water governance.** *Water International.* 2018; **43**(1): 5–12.
Publisher Full Text

24. Freeze RA, Harlan RL: **Blueprint for a physically-based, digitally-simulated hydrologic response model.** *J Hydrol.* 1969; **9**(3): 237–258.
Publisher Full Text

25. Ramamoorthi AS: **Snow-melt run-offstudies using remote sensing data.** *Sadhana.* 1983; **6**(3): 279–286.
Reference Source

26. Idrica: **Blue Dot Observatory**. 2018.
Reference Source

27. Alperen CI, Artigue G, Kurtulus B, *et al.*: **A hydrological digital twin by Artificial Neural Networks for flood simulation in Gardon de Sainte-Croix basin, France.** In: *IOP Conf Ser.: Earth Environ Sci.* IOP Publishing, 2021; **906**(1): 012112.
Publisher Full Text

28. North American National Library of Medicine: **MEDLINE: Description of the Database**. 2022.
Reference Source

29. Institute Jozef Stefan: **IJS Newsfeed: a clean, continuous, real-time aggregated stream of semantically enriched news articles from RSS-enabled sites across the world**. 2022.
Reference Source

Open Peer Review

Current Peer Review Status: ? ✓ ? ?

Version 2

Reviewer Report 09 January 2025

https://doi.org/10.21956/openreseurope.18280.r47777

© **2025 García-Vico Á.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Ángel M. García-Vico
University of Jaén, Jaén, Andalusia, Spain

This paper I have been asked to review presents a data-driven global observatory framework that leverages text mining and complex data visualization to address global challenges, specifically focusing on epidemic intelligence and climate change (water resource management). The framework aggregates data from heterogeneous sources, such as media, scientific publications, and statistical indicators, to deliver actionable insights through real-time monitoring, analysis, and interactive dashboards. The paper in general is well-written and it is very easy to follow. Also, the proposal is interesting for the community. However, I have two main concerns about this work:

1) While the conceptual framework is well-detailed, the technical specifics lack clarity. Can you detail the data cleaning and preprocessing pipeline? How do you handle issues like data bias, missing data, or noise?

2) How does this work compare to other global observatories? For example, how does it outperform or complement existing systems like WHO dashboards or NASA's climate monitoring systems?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computer Science, artificial intelligence

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 29 August 2024

https://doi.org/10.21956/openreseurope.18280.r42254

© 2024 Jemili F. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Farah Jemili
Mars Research Laboratory, ISITCOM, Universite de Sousse, Sousse, Sousse, Tunisia

Strengths:

1. The paper introduces a novel, data-driven global observatory that effectively integrates text mining with complex data visualization to address global challenges. This approach is well-grounded in addressing critical issues like epidemic intelligence and water resource management.
2. The paper meticulously outlines the stages of constructing the global observatory, from data collection to real-time monitoring and visualization, ensuring clarity in how the system operates.
3. The implementation examples, such as the Coronavirus Watch portal and the NAIADES Water Observatory, provide concrete evidence of the system's applicability and effectiveness in real-world scenarios.

Areas for Improvement:

1. Some sections, particularly the methods and implementation, could benefit from clearer language and more concise explanations. Reducing jargon and complex sentences would make the content more accessible.
2. While improvements have been made, the paper could further enhance the discussion on how this work compares with existing solutions, particularly in the areas of text mining and data visualization in global observatories.
3. The paper discusses the system's potential for customization based on local priorities, but more details on how this adaptation is practically implemented would strengthen the argument for its usability and flexibility.

Specific Suggestions:

- **Abstract**: Consider summarizing the key findings more explicitly to highlight the impact of

the proposed observatory.
- **Figures**: Ensure that all figures are clearly labeled and referenced in the text to aid in the reader's understanding.
- **Conclusion**: Expand on the implications of your findings for future research and potential broader applications beyond the discussed domains.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
No

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: AI and Big Data

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 02 January 2024

https://doi.org/10.21956/openreseurope.18280.r36649

© 2024 Leung C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Carson Leung
University of Manitoba, Manitoba, Canada

Costa et al. made some revisions based on previous peer review reports. They described in this 7-page revised submission a data-driven global observatory addressing two worldwide challenges (namely, public health and water) through text mining and complex data visualisation. They

focused on using open data appropriately to monitor news articles and social media feeds related to COVID-19 and water shortage. They would collect, clean, process, integrate and ingest data by making use of key performance indicators (KPI) and digital twin. The presented idea is interesting.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Big data, public health, COVID-19, text mining, social media, data visualisation, data science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 10 July 2023

https://doi.org/10.21956/openreseurope.15624.r32818

© 2023 Leung C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Carson Leung
University of Manitoba, Manitoba, Canada

Costa et al. described in this 8-page submission a data-driven global observatory addressing two worldwide challenges (namely, public health and water) through text mining and complex data

visualisation. They focused on using open data appropriately to monitor news articles and social media feeds related to COVID-19 and water shortage. They would collect, clean, process, integrate and ingest data by making use of KPI and digital twin.

The presented idea is interesting. It would be nice if the authors could consider presenting the work more clearly and providing more details, especially for the corona virus watch observatory. For instance, the authors briefly described how to monitor news articles and social media feeds related to water shortage and quality. However, shorter descriptions were provided for COVID-19.

It would also be nice if the authors could consider providing more details of methods and analysis to allow replication by others. For instance, it is unclear what text mining techniques or what complex data visualisation was used.

Out of the list 26 references, only 5 were from journals. It would be nice if the authors could consider citing some more current formal literature.

It would be nice if the authors could consider further proofreading. For instance, the authors may want to replace the typo "nnd" by "and".

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Big data, public health, COVID-19, text mining, social media, data visualisation, data science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 28 Nov 2023

Joao Costa

We again thank the useful review provided, leading us to improve the content of this paper both in regards to the citations to the related-work and to the details of the study. The time passed between the first submission and this new version also changes the relevance and novelty of the results described, but having in consideration that this is more of a position paper, the content remains useful now that it is better referenced.

Competing Interests: No competing interests were disclosed.

Reviewer Report 29 June 2023

https://doi.org/10.21956/openreseurope.15624.r32819

© **2023 Xu H.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Haowen Xu
Oak Ridge National Laboratory, Bethel Valley Road, USA

Overall Comments:

This manuscript provides a comprehensive review of big data-driven methods for addressing public health and water resources management problems using text mining and complex data visualization. The contributions of this manuscript include some visions and perspectives of media, science, statistics, and sensing over heterogeneous data sources and text mining algorithms, as well as some discussion regarding the implementation of data-driven methods through real-world examples. At the higher level, the methodology and concepts presented in the manuscript align well with the initiative and trend for building smart cities and urban digital twins. The topic is interesting and covers popular research areas in urban informatics. However, this manuscript covers a very wide spectrum of areas (e.g., global water observatory and COVID-19). I recommend the authors refine their scopes and focus. Even within the areas of global water observatory, there are too many perspectives (e.g., hydroinformatics, collaborative watershed management, urban water systems, public engagement, citizen science, and voluntary geographic information). There are also many different types of water resource management (e.g., surface water, groundwater, watershed planning, urban water supply and distribution system, flood and water hazards resilience/vulnerability, river management, and ecohydrology). The "Global Water Observatory" topic is too generic and broad and needs further refinement.

The authors could make this manuscript a review paper or an opinion paper that focuses on an urban data-driven digital twin approach to support urban management and use global water observatories and public health as case studies.

Detailed Comments:

Comment 1:
I think the manuscript only partially cited a portion of current literature. As an example, there are many public data sources for water resources data. I came from an urban computational science and hydroinformatics background in the U.S., where the government has open water data initiatives. Many real-time and historical government sensor data, such as the USGS streamflow, USEPA STORET water quality data, and FEMA Hazus flood data, are open to the public. University consortiums, such as CUAHSI and UCAR, provide online data-sharing and model-sharing platforms (e.g., CUAHSI Hydroshare) to share water resources data and simulations produced in academic research. The authors could include a more comprehensive review of the open data sources for both water resources and COVID data. Is there any government open water data initiatives or University consortiums data initiatives within the EU?

Comment 2:
I would suggest the authors provide a quantitative measure of the trend of different technologies and research areas using Elsevier Scorpous or Google Scholar. These research platforms could provide you with statistics on the literature of a specific research area or contain specific keywords over a few years. A quantified bibliometric analysis is recommended.

Comment 3:
The visualization perspective of the topic is very weak in the current manuscript. There are many review articles that talk about visual computing approaches (information visualization, scientific visualization, and visual analytics) for COVID analysis and water resource management. Please see the example below:

- Leung, C. K., Chen, Y., Hoi, C. S., Shang, S., Wen, Y., & Cuzzocrea, A. (2020, September). Big data visualization and visual analytics of COVID-19 data. In *2020 24th International Conference Information Visualisation (IV)* (pp. 415-420). IEEE.[1]
- Xu, H., Berres, A., Liu, Y., Allen-Dumas, M. R., & Sanyal, J. (2022). An overview of visualization and visual analytics applications in water resources management. *Environmental Modelling & Software*, 105396.[2]

Comment 4:
In Figure 1, the authors mentioned the concept of digital twins. However, an important perspective on digital twins' features is missing from the discussion in the current manuscript. It is the "predictive analytics" capability, which differs from an urban digital twin from previous urban or hydrologic information systems (or decision support systems).

Comment 5:
Section V should provide more examples of digital twins in the water resources management sectors. More cases and real-world implementations should be discussed here. Below is an example; more examples/instances should be reviewed and discussed here.

- Alperen, C. I., Artigue, G., Kurtulus, B., Pistre, S., & Johannes, A. (2021, November). A hydrological digital twin by Artificial Neural Networks for flood simulation in Gardon de Sainte-Croix basin, France. In *IOP Conference Series: Earth and Environmental Science* (Vol. 906, No. 1, p. 012112). IOP Publishing[3].

Comment 6:
The conclusion should also include some technical challenges and aspects of developing urban digital twins for the EU. As an example, what are potential big data or computing challenges from

the cyber-infrastructure or HPC, respectively? Are there any data privacy or data residency concerns across EU countries?

I would recommend a minor revision for this manuscript. Overall the manuscript is in good shape and is well organized.

References

1. Leung CK, Chen Y, Hoi CSH, Shang S, et al.: Big Data Visualization and Visual Analytics of COVID-19 Data. *2020 24th International Conference Information Visualisation (IV), Melbourne, Australia*. 2020. 415-420 Publisher Full Text
2. Xu H, Berres A, Liu Y, Allen-Dumas M, et al.: An overview of visualization and visual analytics applications in water resources management. *Environmental Modelling & Software*. 2022; **153**. Publisher Full Text
3. Alperen C, Artigue G, Kurtulus B, Pistre S, et al.: A Hydrological Digital Twin by Artificial Neural Networks for Flood Simulation in Gardon de Sainte-Croix Basin, France. *IOP Conference Series: Earth and Environmental Science*. 2021; **906** (1). Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Urban Digital Twins, Visual Analytics, Information Visualization, Hydroinformatics, Urban Science, Cyber Physical Systems, Watershed Management (with focus on soil erosion and in-stream sediment transport), Intelligent Transportation System, 3D City Modeling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 28 Nov 2023

Joao Costa

We thank the thorough and useful review. We have added the suggested and other citations, and when possible we have exchanged web articles by research papers. Moreover, in the new submitted version of this paper, we have provided some more content regarding open data and visualisations. We have also improved the readability of the paper.

Competing Interests: No competing interests were disclosed.