Enhancing rare disease research with semantic integration of environmental and health data

Supplementary Information

Albert Navarro-Gallinad[†]
ADAPT Centre for Digital Content, Trinity College Dublin
Dublin, Ireland
albert.navarro@adaptcentre.ie

Fabrizio Orlandi
ADAPT Centre for Digital Content, Trinity College Dublin
Dublin, Ireland
fabrizio.orlandi@adaptcentre.ie

Declan O'Sullivan
ADAPT Centre for Digital Content, Trinity College Dublin
Dublin, Ireland
declan.osullivan@adaptcentre.ie

ABSTRACT

Knowledge Graph (KG) approaches are increasingly being used for data integration processes to combine clinical data with other data sources. Health Data Researchers (HDR) could benefit from these technologies since they require additional types of data outside the health sector, like environmental data, to better understand the extrinsic factors that influence health outcomes in rare disease research. However, using and directly navigating the combined data in the KG can be an obstacle for HDRs. To address this problem, the Semantic Environmental and Rare Disease data Integration Framework (SERDIF) was designed to hide the complexities for these researchers when exploring linked environmental observations with clinical data using a KG approach. The framework was evaluated by HDRs for a case study on Anti-neutrophil cytoplasm antibody (ANCA)-associated vasculitis (AAV) in Ireland, and promising usability and effectiveness results were observed. HDRs studying AAV were able to access, explore and export environmental related data to be used as input for their statistical models. SERDIF has the potential to be a solution for HDRs, who require a flexible methodology to integrate environmental data with longitudinal and geospatial diverse clinical data, in their hypothesis validation of environmental factors for rare disease research.

KEYWORDS

Semantic Data Integration, Knowledge Graph, Usability Evaluation, Environmental Health, Rare Diseases

ACM Reference Format:
Albert Navarro-Gallinad, Fabrizio Orlandi, and Declan O'Sullivan. 2021. Enhancing rare disease research with semantic integration of environmental and health data: Supplementary Information. In *Proceedings of IJCKG '21: The 10th International Joint Conference on Knowledge Graphs (IJCKG '21).* ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

IJCKG '21, December 06–08, 2021, Bangkok, Thailand

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

A SUPPLEMENTARY INFORMATION

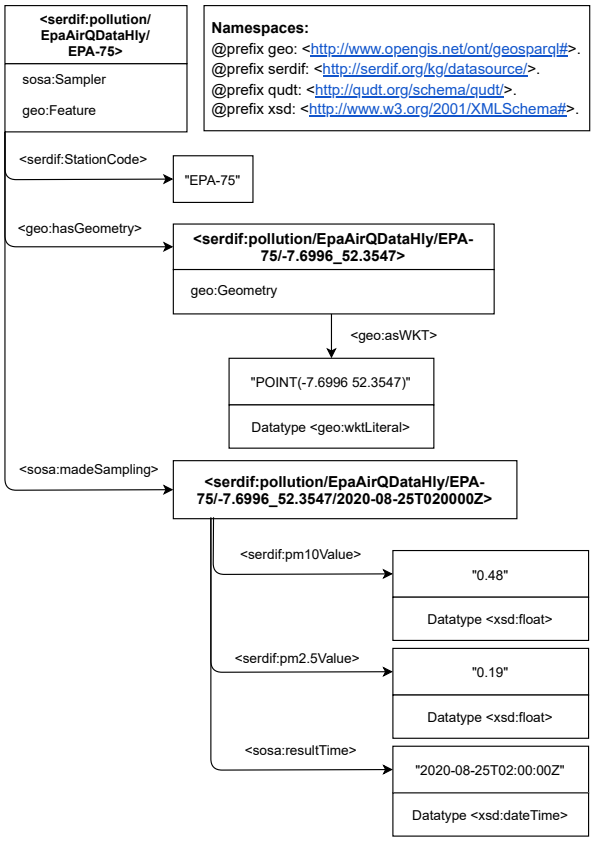

Figure 1: SERDIF sampler data structure diagram.

Table 1: First iteration Thematic Analysis for the AAV in Ireland case study.

Theme	Code	Description	Freq
SERDIF usability	Data visualization is helpful	Data visualizations help to understand the data and flow of the dashboard	32
	Positive user experience	Positive user experience promoted user engagement with linked data	14
	Query features useful	Query features are useful to link and retrieve the required linked data	14
	Data summaries are helpful	Summaries provided a helpful overview of the data for better understanding	12
	Text descriptions and tooltips are handy	Descriptions, tooltips and pop-ups help to guide the user through the processes	9
	Useful	The approach is useful and worthwhile to support for HDR in their research	8
	Easy to use	The tasks are straightforward to complete using the dashboard	7
	Good data exploration features	Data exploration features engage researchers with the linked data	7
	Comparing queries in groups is useful	Comparing queries improves the comprehension of underlying data patterns prior to analysis	7
	Download all queries data at once is practical	Simplified effort to download all data generated during the session	2
Confusing descriptions and features	Clarify text descriptions	Some concepts need further clarification for better understanding	19
	Task instructions clarified	Task wording and structure difficult the acknowledgement of the task completion	12
	Clarify data visualization elements	Some visualizations lacked axis labels and introductory text for a better grasp	10
	Map confusing	Data points density concept and the choropleth map were not practical	8
	Data standardization process not clear	Z-scores aim to help data exploration was misinterpreted	6
	Grouping concept needs clarification	Grouping queries approach was only understood after arranging the groups	6
	Rephrase Ireland selection	Checkbox label to select all LOIs in Ireland had ambiguous meaning	4
Requirements refinement	Clarity data linking process	Data linking and lineage need to be explicit for the participants	25
	Environmental data prior to flare events	Temporal linkage must be for the period before the clinical events	12
	Extend data aggregation options	Sum and median aggregations are requested for the environmental data	5
	Add 'Remission' event category	Compute 'Remission' event category from the existing clinical data	3
	Custom events input	Possibility to import a CSV with events to the KG and visualize it on the dashboard	2
Technical errors	Session technical issues	Delayed responses and control malfunctioning during the session	24
	Grouping queries error	Coding error that did not display certain visualizations in the comparative tab	6