Enhancing Rare Disease Research with Semantic Integration of Environmental and Health Data

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.

CCS CONCEPTS

• **Theory of computation → Data integration**; • **Applied computing** → *Environmental sciences*; • **Human-centered computing** → *Usability testing*;

KEYWORDS

Semantic Data Integration; Knowledge Graph; Environmental Health; Usability Testing; 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. In *The 10th International Joint Conference on Knowledge Graphs (IJCKG'21), December 6–8, 2021, Virtual Event, Thailand.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3502223.3502226

1 INTRODUCTION

Knowledge Graph (KG) approaches emerged as a solution to integrate diverse data sources in response to the existing interoperability challenge in research and industry. In the healthcare domain, the use of KG can facilitate the understanding of diseases leading to identifying new treatments, which could improve patients and citizens quality of life [Esteban-Gil et al. 2017; Kamdar et al. 2019].

This is particularly important when the aetiology and treatment of the disease is unknown as in the vast majority of rare diseases. Applying KG approaches to rare disease research could address the current challenges in grouping similar types of the disease together and linking clinical data with additional types of data outside the health sector such as environmental and social data [Barreto and Rodrigues 2018; Haendel et al. 2020].

However, KGs present a steep learning curve for non-technical researchers [Rietveld and Hoekstra 2017; Smith-Yoshimura 2018]. For example, Health Data Researchers (HDR) exploring the complex questions in environmental health cannot use and navigate themselves clinical and environmental data combined in the KG.

This paper presents the user-centric design and in-use application of the Semantic Environmental and Rare Disease data Integration Framework (SERDIF). The framework aims to support researchers 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. Therefore, the contributions of this paper are in the (i) SERDIF framework: a methodology, the associated knowledge graph structure and a dashboard to provide meaningful access to the linked data. This novel contribution (ii) uses Semantic Web technologies to bridge rare disease research and environmental science disciplines addressing a gap in the state-of-the-art. Furthermore, the paper includes (iii) a description of the framework in-use application for the HELICAL and AVERT projects, which use Semantic Web technologies to link patient and scientific data. SERDIF is envisaged to be used in a series of health data linkage projects at a European and international level leading to possibly hundreds of interested HDRs.

The paper is structured as follows. Section 2 presents the SERDIF general approach. Section 3 describes the application of SERDIF for HDRs studying ANCA-Associated Vasculitis (AAV) in Ireland. Section 4 reports and discusses the usability evaluation of the framework. Section 5 reviews related work. Section 6 describes the impact and uptake of this research.

IJCKG'21, December 6–8, 2021, Virtual Event, Thailand

ACM ISBN 978-1-4503-9565-6/21/12…$15.00
https://doi.org/10.1145/3502223.3502226

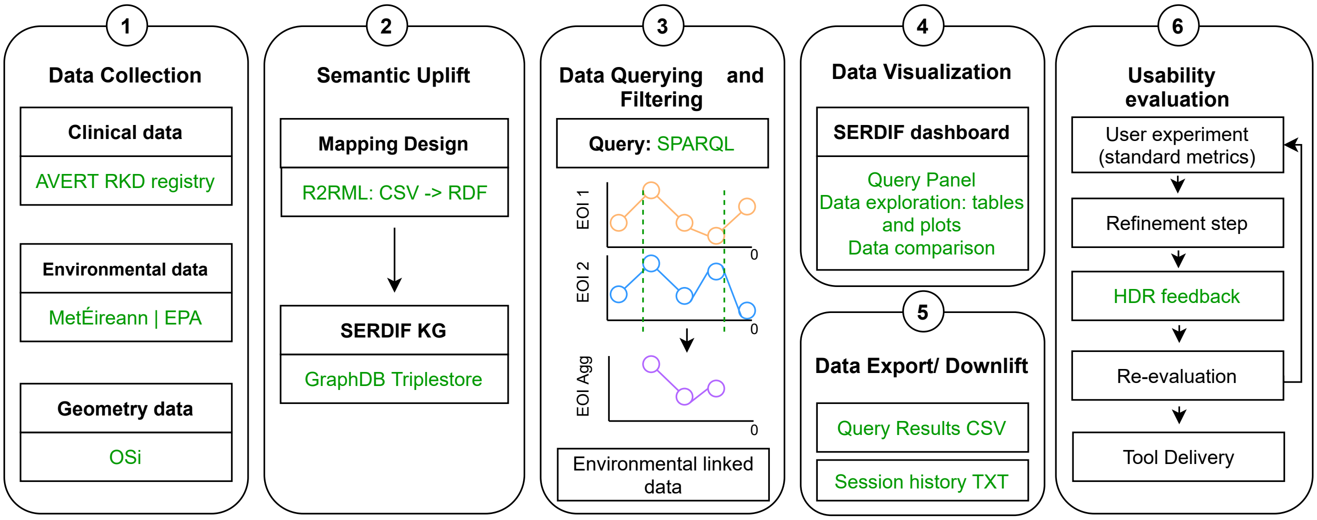


Figure 1: SERDIF approach diagram with the in-use application in green.

2 THE SERDIF APPROACH

SERDIF was developed as a result of the state of the art review presented in this paper (see *Section 5*) and the outcome of the usability evaluation conducted on an initial dashboard [Navarro-Gallinad et al. 2020]. The framework is a combination of a methodology, a knowledge graph and a dashboard.

The **methodology** is a series of steps that should be taken to define the necessary spatio-temporal data structures to combine clinical and environmental data. The methodology is divided into six main processes illustrated in Fig. 1.

(1) Data collection: this process requires accessing existing clinical data and downloading environmental and geometry data. Clinical data comprises any data type format with temporal and spatial components which are interpreted as geolocated events. Environmental data consists of observation data represented as geolocated time series. Geometry data include the necessary region geometries containing the locations from the clinical and environmental data.

(2) Semantic uplift: this process designs a declarative mapping to uplift the data gathered from the data collection process to RDF [Brickley and Guha 2014]. The geometries used in the mapping must be GeoSPARQL types (point, line, polygon, multipolygon, etc.) for the downlifting section to reason over the spatial dimension of the data [Geo 2012]. Furthermore, this process includes the conversion of relational or tabular data to RDF adding semantics. Engines like R2RML [Das et al. 2012] offer the framework the ability to convert those files using a mapping, generating RDF for the KG from a table or relational database. The semantic uplift process is completed with the RDF graphs being uploaded to a triplestore that supports GeoSPARQL.

(3) Data querying and filtering: this process defines a spatio-temporal query as a SPARQL template. A SPARQL query template that has placeholders (or variables) for users' input (see next step) and it is designed to be generic enough to adapt to different data sources. The linking between environmental and clinical data occurs during the SPARQL query reasoning over location and time.

(4) Data visualization: this process designs an initial visual tool to grant meaningful access to domain experts hiding the complexities in using Semantic Web technologies. The tool design is user-centric, focused on domain experts' requirements, to develop an effective tool. The initial requirements can be extracted from expert consensus within a project.

(5) Data exporting/downlift: this process exports combined and/or aggregated data from the Knowledge Graph in tabular format for analysis. The results from the SPARQL query can be exported as a table (CSV), which typically is on of the preferred input formats for data analysis. The results can also be exported in other data formats like JSON if required. A log from the queries should also be stored in text format with the selected query input options in case the user wants to recover previous queries.

(6) Usability evaluation: this process starts with the evaluation of the visual tool. Standard evaluation metrics are required for this step, enabling comparison of prototype tools with later versions of the tool, as well as with other tools. The combination of different metrics provides more information to assess the achievement of the user requirements for the tool to be effective. Following, this process refines the requirements and framework artefacts based on the evaluation outcome. The outcome is used to improve the usability and effectiveness of the methodology, knowledge graph and tool by updating the existing version. The usability evaluation is conducted in an iterative manner until agreement is reached with users in fulfilling the requirements. Once the users are

satisfied, the visual tool (i.e. dashboard) will be ready to be delivered.

Therefore, the methodology facilitates the integration of diverse data sources through the use of a KG and guarantees the scalability of new data sources being added. The open-ended approach presents an advantage in rare disease research accounting for a possible increase in the data volume and sources, which is likely to change as research progresses and more knowledge is gained in understanding the complex environmental-patient system.

The **knowledge graph** benefits from the spatio-temporal data structures to combine clinical and environmental observations through locations, from geometry data; and relative periods from the clinical events.

The **dashboard** is designed from a user-centric perspective to support HDRs access, explore and export the linked spatio-temporal environmental data by aiding a HDR formulate a query in an intelligible non-technical manner and to explore the data with appropriate visualizations.

3 SERDIF IN-USE

The HEalth data LInkage for ClinicAL benefit (HELICAL)[1] is an European project with the goal of finding solutions to the challenges faced by patients with rare diseases when it comes to connecting their personal data with scientific data. In this project, Semantic Web technologies are the approach chosen to combine multiple diverse data sources with spatial and temporal common features as medical registries and environmental data.

One of the studied rare diseases is AAV, a rare autoimmune disease of unknown aetiology which affects small blood vessels in different parts of the body in a progressive manner, resulting in damage to vital organs. The current theory sustains that this aetiology involves a complex interaction between environmental and epigenetic factors, in a genetically susceptible individual [Kitching et al. 2020]. The suspicion of an environmental trigger emerges from the spatiotemporal clustering of the disease, supported by the seasonality, latitudinal gradient in disease onset and the urban/rural prevalence [Scott et al. 2020]. Understanding the environmental trigger could lead to predicting when flares of the disease may occur for individual patients.

While interoperable disease registries combined with environmental data could facilitate this research, knowledge engineers are required in the process to perform the queries to fulfil the researchers needs. The intention going forward in similar healthcare data linkage projects is to allow the researchers themselves to access, explore and retrieve the clinical and environmental data represented and linked through Semantic Web technologies.

Therefore, the AAV paradigm is an ideal opportunity to apply the SERDIF framework to enable hypothesis validation of the environmental triggers for this disease. This case study allows the framework to be evaluated in a real situation supporting HDRs meaningful access to a variety of linked data sources, clinical and environmental, which have been carefully combined. The following items refer to the instantiation of the SERDIF methodology for AAV in Ireland outlined in the diagram from Fig. 1. An example of the framework for this use case is made open source and accessible through GitHub for reproducibility of the paper (i.e. with synthetic clinical data and reduced environmental data).

https://github.com/navarral/ijckg2021-serdif-paper

3.1 Data collection process

Clinical, environmental and geometry data are manually collected (or accessed in the case of clinical data) in the data collection process. Previous work from the AVERT project[2] facilitated the access of clinical data, which were already uplifted to RDF [Reddy et al. 2019]. The events described in the clinical data are AAV patient flares geolocated in an electoral district or hospital within the Republic of Ireland. Consequently, geometries of all the counties in the Republic of Ireland are gathered from the OSi resource as RDF files[3].

The interest from HDRs is the validation of environmental triggers for AAV; therefore, environmental data is gathered from land-based stations within the country. In the first iteration, weather[4] and pollution[5] data are collected as tabular files (CSV).

In addition, metadata files that include the environmental variables descriptions and station locations for each data source are also gathered.

3.2 Semantic uplift process

The semantic uplift process provides an R2RML mapping specific to each environmental data source (i.e. including the specific variables in the triple maps) but keeping the same data structure for metadata and data files across sources. The data structure re-uses the Sensor Network (SOSA) existing vocabulary[7] facilitating spatiotemporal reasoning due to the organization levels (see SI on Github for a snippet of the KG): geolocated samplers (sosa:Sampler) that include samplings (sosa:Samplings) as time series data. Observation values are described using a custom approach (e.g. serdif:SO2value) since no appropriate environmental vocabulary with this description had been identified. In addition, the sampler's location is modelled with GeoSPARQL and the time series as xsd:dateTime, enabling the spatial and temporal reasoning in the following step. The data structure proposed in this research is based on the initial requirements gathered from expert consensus [Navarro-Gallinad et al. 2020].

Regarding the implementation, R2RML-F is the R2RML engine used in this step allowing access to CSV files as relational tables [Debruyne and O'Sullivan 2016] in the uplift process. Furthermore, this engine has a functionality of using transformation functions for data from the CSV files which is used to convert the raw date time syntax to the adequate standard syntax for RDF files.

The environmental RDF files generated together with the clinical and geometry graphs are imported to a GraphDB triplestore[8] (see paper's GitHub). Table 1 summarizes the data graphs imported into the triplestore in terms of data type, access, provenance format, temporal granularity (spatial for geometry data) and the number of triples per graph. Importing the RDF graphs includes a validation step that checks for any syntax errors which stops on error. The

[1] http://helical-itn.eu/

[2] https://www.tcd.ie/medicine/thkc/avert/

[3] http://data.geohive.ie/

[4] https://www.met.ie//climate/available-data/historical-data

[5] https://airquality.ie/

[7] www.w3.org/ns/sosa/

[8] https://graphdb.ontotext.com/

Table 1: Data sources summary for the Republic of Ireland during 2013-2020 period.

Data source	Data type	Access	Format	Granularity	# Triples
Clinical	Disease registry	Private	RDF	Daily	1.4M
Weather	Land-based station	Public	CSV	Hourly	27.8M
Pollution	Land-based station	Public	CSV	Hourly	2.5M
Geometry	Multipolygons	Public	RDF	County/ED[6]	28k

triplestore was chosen due to the GeoSPARQL support, key for HDR queries, and their easy to use interface to develop applications.

3.3 Data querying and filtering process

In this process, the clinical and environmental data sets could have be linked using different approaches: building an ontology, making sure the same URI is shared for both datasets (e.g. as manual input in the mappings) or using a SPARQL query. The SPARQL query linkage method is recommended because of the possibility to reason over location and time at a query level (see Fig. 2).

Location. The RDF clinical graph includes patient electoral district and hospital location compliant with GeoSPARQL geometries (?eventGeom), polygon and point respectively; and the environmental graph contains point locations for the land-based measurement stations (?envoGeom). Therefore, reasoning is necessary due to the missing explicit triple pattern shared between both data sources (i.e. the point geometries do not concur). GeoSPARQL functions [Geo 2012] enable spatial reasoning between geometries with functions like geof:distance or geof:sfWithin. In this case, geof:sfWithin is the function chosen since the aim is to aggregate environmental observations within a region (?regionGeom), and then associate the aggregation with an individual patient record within the same region.

Time. Individual patient records contain events such as disease activity and remission state dates or hospital admissions represented as xsd:dateTime datatypes (?dateEvent). Hence, environmental observations are filtered for a specific period related to the clinical events. In this case, the period is defined by the lag from the event (?dateLag) and the duration (?dateStart), which in Fig. 2 is of 7 and 30 days respectively.

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

# Spatial reasoning
FILTER(geof:sfWithin(?eventGeom, ?regionGeom))
FILTER(geof:sfWithin(?envoGeom, ?regionGeom))

# Temporal reasoning
BIND(?dateEvent - "P7D"^^xsd:duration AS ?dateLag)
BIND(?dateLag - "P30D"^^xsd:duration AS ?dateStart)
# Filter environmental data for the selected dates
FILTER(?dateObs > ?dateStart && ?dateObs <= ?dateLag)
```

Figure 2: Spatiotemporal reasoning used in the SERDIF querying process as a SPARQL example.

3.4 Data visualization

An initial dashboard has been designed to hide the complexities in executing queries against a triplestore for Health Data Researchers (HDR), and provides further comprehension of the environmental linked data with summaries, data tables and plots. *Dash*[9] is the Python framework used to build the dashboard, which contains query input and main panels in a coordinated view. For a better understanding of the data visualization step an example dashboard is made available at:

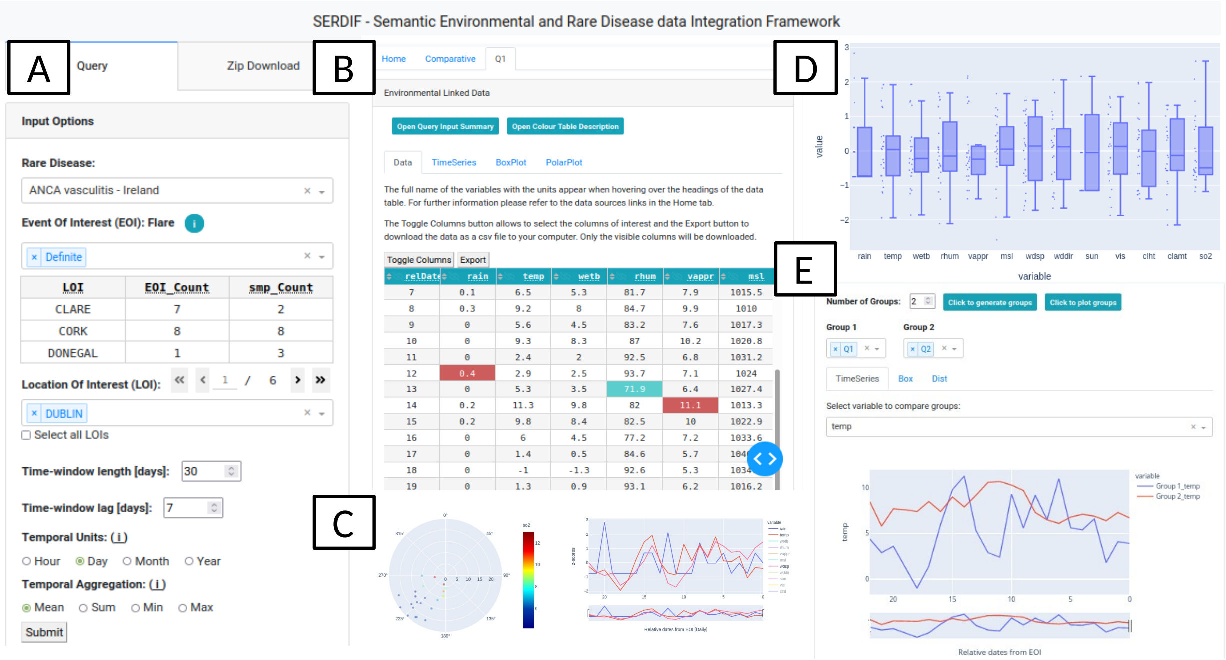
https://serdif.adaptcentre.ie/dashboard

Query input panel: the user can select multiple options as input from the clinical data to retrieve the associated environmental data in this panel (see Fig. 3 A). The input options are dynamically displayed using live predefined queries executed while the dashboard is running, providing a flexible visual tool. If new data becomes available as the research progresses, the dashboard will be able to adapt to the new data automatically. The query inputs are also sequential providing a data validation step per selected input (i.e. the following option does not become available if the previous is missing or not selected). A final SPARQL ASK query enables the submit button at the bottom of this panel when all the options are selected and data is available. When the user clicks on the submit button, the selected options are substituted into the SPARQL query template from the previous step, URL encoded and executed against the data in the triplestore.

Main panel: the main panel of the dashboard consists of three tabs: home, comparative and query number (see Fig. 3 B). The home tab provides an introduction to the dashboard use together with acknowledgement of the data sources combined throughout their website links. In addition, an interactive choropleth map is available to explore the number of samplers per county.

After each query submission, a new tab is generated with a summary of the input selections and four sub tabs named (i) data, (ii) time series, (iii) box and (iv) polar plot. The (i) data sub tab displays the raw data outcome resulting from the query as a heat map data table. The plot sub tabs (see Fig. 3 C and D) are interactive and allow the user to visually explore the data table normalized variables (ii) to identify any internal structure (i.e. autocorrelation, trends or seasonality); and (iii) to study the variability and distribution per variable and relation to the other variables which can contribute to further understanding the environmental trigger. (iv) The polar plot facilitates comprehension of the complex relationship between environmental variables and wind.

[9] https://plotly.com/dash/



SERDIF - Semantic Environmental and Rare Disease data Integration Framework

Query | Zip Download

Input Options

Rare Disease: ANCA vasculitis - Ireland

Event Of Interest (EOI): Flare

Definite

LOI	EOI_Count	smp_Count
CLARE	7	2
CORK	8	8
DONEGAL	1	3

1 / 6

Location Of Interest (LOI): DUBLIN

□ Select all LOIs

Time-window length [days]: 30

Time-window lag [days]: 7

Temporal Units: (i)

○ Hour ● Day ○ Month ○ Year

Temporal Aggregation: (i)

● Mean ○ Sum ○ Min ○ Max

Submit

Home | Comparative | Q1

Environmental Linked Data

Open Query Input Summary | Open Colour Table Description

Data | TimeSeries | BoxPlot | PolarPlot

The full name of the variables with the units appear when hovering over the headings of the data table. For further information please refer to the data sources links in the Home tab.

The Toggle Columns button allows to select the columns of interest and the Export button to download the data as a csv file to your computer. Only the visible columns will be downloaded.

Toggle Columns | Export

relDate	rain	temp	wetb	rhum	vappr	msl
7	0.1	6.5	5.3	81.7	7.9	1015.5
8	0.3	9.2	8	84.7	9.9	1010
9	0	5.6	4.5	83.2	7.6	1017.3
10	0	9.3	8.3	87	10.2	1020.8
11	0	2.4	2	92.5	6.8	1031.2
12	0.4	2.9	2.5	93.7	7.1	1024
13	0	5.3	3.5	71.9	6.4	1027.4
14	0.2	11.3	9.8	82	11.1	1013.3
15	0.2	9.8	8.4	82.5	10	1022.9
16	0	6	4.5	77.2	7.2	1033.6
17	0	1.4	0.5	84.6	5.7	[illegible]
18	0	-1	-1.3	92.6	5.3	1034
19	0	1.3	0.9	93.1	6.2	1016.2

Number of Groups: 2

Click to generate groups | Click to plot groups

Group 1: Q1

Group 2: Q2

TimeSeries | Box | Dist

Select variable to compare groups: temp

Figure 3: Screenshot of the SERDIF dashboard displaying (A) the query input panel, which allows non-Semantic Web experts to access the knowledge graph; (B) the tab generated after submitting a query, which includes a query summary, a data table and three different visualizations, (C) polar, time series and (D) box plots; and (E) the comparison tab, where previous queries can be compared in groups.

Previous queries can be visually compared in the comparative tab by specific variable (see Fig. 3 E). In addition, the queries can be arranged into groups to potentially reveal signals that the individual queries were hiding.

3.5 Data export/downlift process

The query number tab generated after each query submission contains an export button situated on top of the data table (see Fig. 3 B). The user can click the export button to download the selected columns from the data table as a CSV file. Moreover, the user can also export all data tables resulting from previous queries during the session as a ZIP file. The feature to download data tables as a zipped file is located in the ZIP Download tab from the Query input panel (see Fig. 3 A).

4 EVALUATION

The first iteration of the SERDIF dashboard has been evaluated using HDRs from the Irish AAV case study. The interaction and evaluation with these domain experts has been conducted through the usability study described below.

Experimental Setup and Execution. The sample size for the AAV usability study is of 10 HDRs without practical experience in Semantic Web technologies. The researchers are international professors, researchers and PhD students with fluent English, who are analysing AAV clinical data in their research. The sample size of 10 covers the requirements of a specialised tool for validation of clinical and linked environmental data [Macefield 2009].

Participants were asked to complete seven tasks, designed to assess the three core requirements, following a concurrent think-aloud protocol (CTA) [Boren and Ramey 2000] (i.e.listening to the participant process while completing the tasks). The tasks were derived from consensus among HDRs with real workflows in mind. The overview of the tasks were: reading and understanding the 'Home' tab (T1), submitting a query (T2), comprehending data table with the query results (T3), exploring the results with the available plots (T4), submit two more queries and compare them (T5), exporting the results as a csv and zip (T6) and summarizing the overall experience in completing the tasks with the SERDIF dashboard (T7). As the experiment was conducted during COVID-19 restrictions, synchronous remote testing was the method used through a video conferencing platform with remote control functions.

The participants think-aloud statements and extra feedback were recorded by hand, supported by the meeting automatic transcripts. The text statements and transcripts were analysed qualitatively with Thematic analysis, following the six-step process from Nowell et al. [Nowell et al. 2017] The steps involve familiarization with the collected data, generating initial codes, searching for themes, reviewing the themes, defining and naming themes and producing the report.

The post-questionnaire used in this experiment is the the Post-Study System Usability Questionnaire (PSSUQ), which is a standard questionnaire meant to assess the usability evolution during the development of a system with 19 questions (second version of the questionnaire was used in this study) [Lewis 2002].

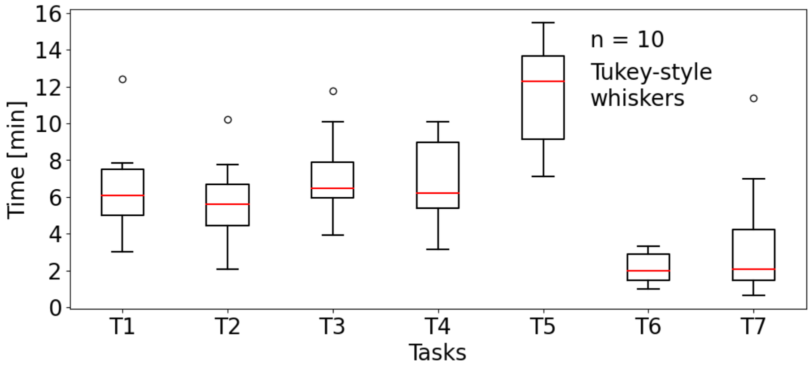


Figure 4: Time spent to complete each task during the usability experiment.

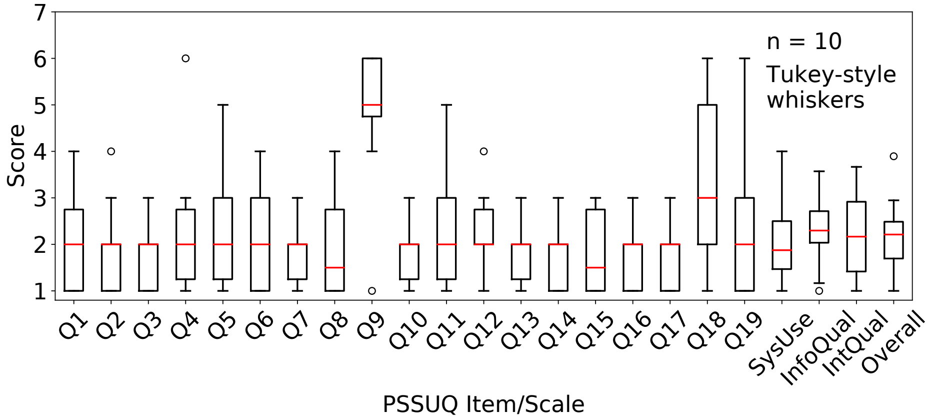


Figure 5: PSSUQ scores box plot with the four averaged metrics (SysUse, InfoQual, IntQual and Overall) on the right end with a sample sizes of 80, 70, 30 and 190. The scores are in a Likert 7 points scale where the lower the value the higher the satisfaction.

The methods described above include successful completion of the tasks following a CTA protocol and time on task to support the PSSUQ standard questionnaire. The CTA protocol grants feedback to understand the effective task completion and time on task, in a meaningful way. The methods combine quantitative and qualitative metrics to evaluate the usability of the SERDIF framework through the dashboard artefact for HDRs.

The quantitative results of this experiment include the time per task (Fig. 4) and the PSSUQ scores (Fig. 5).

Time per task. The box plots in Fig. 4 compare participants' time spent on tasks, which all the participants completed successfully. However, one of the steps in task 5 could not be completed for a couple of participants due to their query selections and a coding error which affected the visualization of the comparative plots. Tasks were categorized into four complexity groups based on the length of the boxes (IQR) and the median time per task: low (IQR < 3min and Median < 3min; T6, T7), moderate (IQR < 3min and Median < 7min; T1, T2), high (IQR > 3min and Median < 7min; T3, T4) and very high (IQR > 3min and Median > 7min; T5). Therefore, the tasks seem to follow an increasing pattern of complexity from T1-T5 and drop for the last two tasks, T6 and T7.

PSSUQ: The Post-Study System Usability Questionnaire. The second type of quantitative results are the PSSUQ scores and scale, which follow a 7-point Likert Scale and assesses four different metrics: system usefulness (SysUse), information quality (InfoQual), interface quality (IntQual) and overall, averaged from 1-8, 9-15, 16-18, 1-19 questions. In this scale, the lower the value, the higher the satisfaction. Most of the box plots in Fig. 5 for the PSSUQ scores have a median of 2 (15 out of 19 questions) and a spread below 2 points (16 out of 19 questions). Error messages (Q9), expected capabilities and functions (Q18) and overall satisfaction (Q19) are the worse scores per question; becoming productive (Q8) and information organization (Q15) are better.

The qualitative results are gathered from the notes and transcriptions taken during the experiment sessions which followed the CTA protocol. The text transcriptions were analysed following the six-step process Thematic analysis previously outlined. First, the notes and transcripts were read to familiarize the researcher with the data. Then, the transcripts were annotated with appropriate codes using the QualCoder[10] software to facilitate this process. The codes were identified following an inductive approach without trying to

[10] https://github.com/ccbogel/QualCoder/releases/tag/2.4

Table 2: First iteration Thematic Analysis summary for the AAV in Ireland case study.

Themes	Code Description summary	Total Frequency
SERDIF dashboard Usability	Positive overall user experience emphasizing the data exploration features and the usefulness of SERDIF	112
Clarify description and features	Some complicated jargon and ambiguous text descriptions	65
Requirements refinement	Unclear data lineage and environmental data linked to a period prior to the flare events	46
Technical errors	Delays and control malfunctioning during the virtual experiment session	30

fit a pre-existing coding frame. The codes were grouped by similarity to identify patterns and themes, which were then reviewed to support the theme selection. The themes were identified following the semantic approach with the explicit or surface meanings of the transcription. The resulting themes were named and defined based on the code descriptions within each theme and then reported as a table (see Table 2). This table includes the code descriptions and the frequency of the code in the transcripts (see paper's GitHub) for a more detailed view).

Table 2 summarizes the codes description and frequency for the first iteration of the AAV in Ireland case study. The emerging themes from the Thematic analysis were (1) SERDIF dashboard usability, (2) requirements refinement, (3) clarify descriptions and features and (4) technical errors. The themes identified support the initial hypothesis of SERDIF dashboard being an adequate approach to support HDR, while guiding the development towards an effective tool by refining the requirements and improving the descriptions. Furthermore, the themes acknowledged the perception associated with the complex topic of comprehending linked environmental data to support rare disease research and the importance of findings across multiple patients.

The combination of the quantitative and qualitative results were analysed towards the assessment of the three initial requirements from Navarro-Gallinad et al. 2020 [Navarro-Gallinad et al. 2020]. In the next step of the usability evaluation process, the requirements and the tool are refined according to the results obtained from the first iteration of the evaluation with the goal of increasing the usability for HDRs.

Requirements elicitation and analysis. The requirements are refined following an iterative process based on the usability evaluation results. During the process constant feedback from partners and HDRs is incorporated. The main requirements collected after the first iteration resulted as:

Requirement 1: Enable HDR to query specific clinical data to retrieve linked environmental data from a defined period prior to multiple patients' flare events within a region, without the need for knowledge of the underpinning semantic web technologies.

Requirement 2: Support the understanding of the HDR in the use, limitations and data lineage of the linked environmental data to support identification of flares for rare diseases.

Requirement 3: Allow for the download of selected linked environmental data to be used as input in statistical models for data analysis.

The first iteration of the SERDIF framework evaluated with the AAV in Ireland case study yielded a satisfactory outcome. First, the methodology to integrate environmental data with longitudinal and geospatial diverse clinical data proved to be useful for HDRs for this first case study. Second, the associated knowledge graph structure resulted in effectively linking graph data through a SPARQL query. Third, the SERDIF dashboard allowed researchers to access, explore and export to linked environmental data.

However, SERDIF needs to improve the data lineage transparency in the combining process, as well as, clarifying the text descriptions in the dashboard for the domain matter experts. SERDIF data lineage transparency will have to be assessed to comply with the General Data Protection Regulation law (GDPR)[11] for processing individual environmental-patient linked records in the next iteration. Due to the nature of rare disease data, the environmental-patient linked data cannot be assumed to be completely anonymized, hence pseudoanonymized data, presenting a data protection risk for the patients when combined with other data sources, such as social media or statistical data. Therefore, the following iterations will include a data protection step complying with GDPR. The KG will be edited accordingly to support this new process.

Furthermore, an alternative method to conduct the usability experiment will be explored to address the delay and control malfunctioning of the current remote control functionality of the video conferencing platform. The following iteration of the framework will be conducted with the refined requirements and framework artefacts on the same use case, AAV in Ireland.

When agreement is reached within the AAV in Ireland case study, the intention is to further validate SERDIF with Kawasaki disease in Japan case study with the same usability evaluation. Climatological studies point towards an environmental agent transported by tropospheric winds to be the trigger link of this paediatric vasculitis [Rodó et al. 2014]. For that reason, it is an ideal scenario to apply SERDIF, supporting HDRs in their hypothesis validation.

5 RELATED WORK

This section overviews the state-of-the-art in combining methods for rare disease clinical data with other data sources using Semantic Web technologies. The review is centred on: the clinical data type, other data sources integrated, methodology steps covered compared to SERDIF (see *Section 2*), the requirement of users to be Semantic

[11] https://gdpr-info.eu/

Web experts to use the methods and tools and whether a usability test to evaluate the effectiveness and usefulness of the approach has been undertaken.

Roos et al. discuss the impact of applying Semantic Web technologies to answer rare disease research questions [Roos et al. 2017]. The application requires collaboration between domain experts and computer scientists to address the methodological, representational and automation challenges for correctly combining data from the dispersed resources. The researchers could collaborate by designing a visual interface from a domain expert requirements that benefits from the underlying technologies for an easy and meaningful access to the combined data, as proposed in the SERDIF approach.

Visual interfaces have also been used in the biomedical area of rare disease research to facilitate the access to linked data. For example, interfaces granting secure-access for clinical researchers that operate on top of linked international biobanks and registries as the RD-Connect platform [Gainotti et al. 2018; Thompson et al. 2014]. The platform joined in a collaborative approach NeurOmics and EURenOmics to advance forward the -omics research and data sharing for the Rare Disease community, indicating the importance of this type of research [Lochmüller et al. 2018]. Following, the DisGeNET [Pinero et al. 2015] and LORD platforms include a web interface, allowing the user to perform free-text searches from a gene- or disease-centric view of the data to answer questions related to rare diseases (DisGeNET), and to navigate through the relationships between rare diseases supporting health information systems routines (LORD). However, the reviewed platforms above did not include a usability evaluation to assess the usefulness of the visual interface for domain experts.

In the same biomedical domain, other approaches require users with Semantic Web practical expertise (e.g. building a query) to benefit from the interfaces. This is the case for Mina et al. [Mina et al. 2015] and the SCALEUS visual interface [Sernadela et al. 2017a,b], which combined genetic and epigenetic data sources for Huntington disease research and Electronic Health Records (EHR) from individual patients with genetic data, respectively. The latter is the only reviewed study that conducts a usability test, following a customized approach to evaluate the visual interface impeding the comparison with similar visualization tools.

In contrast to the related work outlined above, our research presents a framework (SERDIF) that includes a usability evaluation process with standard metrics. The usefulness of the dashboard artefact is evaluated for Health Data Researchers granting validation of the underlying methodology and Knowledge Graph. The standard metrics will allow comparisons with similar tools, which could not have been performed from the reviewed studies. Furthermore, our research carefully combines disease registries with environmental observations providing a new type of data source to the predominant focus in biomedical data.

6 FUTURE WORK AND CONCLUSIONS

The volume of clinical and environmental data is expected to increase as research progresses and more stakeholders become interested in using SERDIF. The framework will have to be capable of working with different clinical data inputs such as administrative data, disease registries and health surveys. In the upcoming case study, national survey data reporting daily cases of Kawasaki disease for the 47 prefectures in Japan. Furthermore, other queries to explore the complex environmental-patient system might become relevant and new complex relationships (e.g. influence on patient treatment) will have to be considered.

Nevertheless, the framework is not only limited to rare diseases but could be applied to other diseases as respiratory and cardiovascular diseases (e.g. COVID-19), where the integration of environmental data could lead to gain new biological insights. The requirements to apply SERDIF are that the disease events are geolocated with an existing geometry (e.g. a lat/lon point, region, area or country code) and made accessible, even an example with anonymized data will suffice.

Besides the novelty of SERDIF in the Semantic Web community described in *Section 5*, the framework has the potential to become a component of an early warning system for public health management Subject Matter Experts (SME). The warning system could benefit from the existing knowledge graph through the triplestore API and provide the required data to assess the patient's risk in having a relapse. The risk assessment would influence the monitoring and current treatment to improve patients life quality. The APIs to support such system and the quality assurance to be a trusted component within a decision support architecture will be assessed in the next versions.

SERDIF facilitates the work of researchers (i.e. hiding the complexities for non-technical users) who in turn will be more effective in finding new patterns and solutions/treatments for citizens. For example, promoting studies of the environmental factors for attendees to several venues like schools or factories could possibly benefit from the framework. Furthermore, the framework could possibly impact other domains like climate change, environmental studies, biology or even the food industry enlightening the environmental effects on human health.

From the design and the results obtained from the evaluations performed to date, there are indications that SERDIF will yield a satisfactory outcome for the first rare disease case study, AAV in Ireland. Researchers reported a positive and useful experience when using Knowledge Graphs through the SERDIF dashboard to address their current data integration challenges. Researchers from AVERT and HELICAL projects (n≈30) showed interest in using the framework to validate their hypothesis to address complex environmental health research questions.

In conclusion, the impact of SERDIF will range from enabling HDRs to validate their hypothesis of environmental factors in rare disease research, to influencing policy making for climate change and data protection risks.

ACKNOWLEDGMENTS

This research was conducted with the financial support of HELICAL as part of the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 813545 and No. 713567 at the ADAPT Centre for Digital Content Technology (grant number 13/RC/2106_P2) at Trinity College Dublin.

REFERENCES

2012. GeoSPARQL - Semantic Web Standards. https://www.w3.org/2001/sw/wiki/GeoSPARQL. (Sept. 2012).

Mauricio L. Barreto and Laura C. Rodrigues. 2018. Linkage of Administrative Datasets: Enhancing Longitudinal Epidemiological Studies in the Era of “Big Data”. 5 (Dec. 2018), 317–320.

T. Boren and J. Ramey. 2000. Thinking Aloud: Reconciling Theory and Practice. 43 (Sept. 2000), 261–278.

Dan Brickley and R.V. Guha. 2014. Resource Description Framework (RDF) Model and Syntax Specification. https://www.w3.org/TR/rdf-schema/. (Feb. 2014).

Souripriya Das, Seema Sundara, and Richard Atkinson. 2012. R2RML: RDB to RDF Mapping Language. https://www.w3.org/TR/r2rml/. (Sept. 2012).

Christophe Debruyne and Declan O’Sullivan. 2016. R2RML-F: Towards Sharing and Executing Domain Logic in R2RML Mappings. *Proceedings of the Workshop on Linked Data on the Web, LDOW2016, co-located with the 25th International World Wide Web Conference (WWW 2016), Montreal* (2016), 5.

Angel Esteban-Gil, Jesualdo Tomás Fernández-Breis, and Martin Boeker. 2017. Analysis and Visualization of Disease Courses in a Semantically-Enabled Cancer Registry. 8 (Sept. 2017), 46.

Sabina Gainotti, Paola Torreri, and Chiuhui Mary et al. Wang. 2018. The RD-Connect Registry & Biobank Finder: A Tool for Sharing Aggregated Data and Metadata among Rare Disease Researchers. 26 (May 2018), 631–643.

Melissa Haendel, Nicole Vasilevsky, and Deepak et al. Unni. 2020. How Many Rare Diseases Are There? 19 (Feb. 2020), 77–78.

Maulik R. Kamdar, Javier D. Fernández, and Axel et al. Polleres. 2019. Enabling Web-Scale Data Integration in Biomedicine through Linked Open Data. 2 (Dec. 2019), 90.

A. Richard Kitching, Hans-Joachim Anders, and Neil et al. Basu. 2020. ANCA-Associated Vasculitis. 6 (Aug. 2020), 1–27.

James Lewis. 2002. Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. 14 (Sept. 2002), 463–488.

Hanns Lochmüller, Dorota M. Badowska, and Rachel et al. Thompson. 2018. RD-Connect, NeurOmics and EURenOmics: Collaborative European Initiative for Rare Diseases. 26 (June 2018), 778–785.

Ritch Macefield. 2009. How To Specify the Participant Group Size for Usability Studies: A Practitioner’s Guide. *Journal of usability studies* 5, 1 (2009), 12.

E. Mina, M. Thompson, and K. M. et al. Hettne. 2015. Multidisciplinary Collaboration to Facilitate Hypotheses Generation in Huntington’s Disease. In *2015 IEEE 11th International Conference on E-Science*. 118–125.

Albert Navarro-Gallinad, Alan Meehan, and Declan O’Sullivan. 2020. The Semantic Combining for Exploration of Environmental and Disease Data Dashboard for Clinician Researchers. In *Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data (VOILA!2020), co-located with the ISWC2020*.

Lorelli S. Nowell, Jill M. Norris, and Deborah E. et al. White. 2017. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. 16 (Dec. 2017), 160940691773384.

J. Pinero, N. Queralt-Rosinach, and A. et al. Bravo. 2015. DisGeNET: A Discovery Platform for the Dynamical Exploration of Human Diseases and Their Genes. 2015 (April 2015), bav028–bav028.

Brian P Reddy, Brett Houlding, and Lucy et al. Hederman. 2019. Data Linkage in Medical Science Using the Resource Description Framework: The AVERT Model. 1 (March 2019), 20.

Laurens Rietveld and Rinke Hoekstra. 2017. The YASGUI Family of SPARQL Clients 1. 8 (Jan. 2017), 373–383.

Xavier Rodó, Roger Curcoll, and Marguerite et al. Robinson. 2014. Tropospheric Winds from Northeastern China Carry the Etiologic Agent of Kawasaki Disease from Its Source to Japan. 111 (June 2014), 7952–7957.

Marco Roos, Estrella López Martin, and Mark D. Wilkinson. 2017. Preparing Data at the Source to Foster Interoperability across Rare Disease Resources. In *Rare Diseases Epidemiology: Update and Overview*, Manuel Posada de la Paz, Domenica Taruscio, and Stephen C. Groft (Eds.). Springer International Publishing, Cham, 165–179.

Jennifer Scott, Jack Hartnett, and David et al. Mockler. 2020. Environmental Risk Factors Associated with ANCA Associated Vasculitis: A Systematic Mapping Review. 19 (Nov. 2020), 102660.

Pedro Sernadela, Lorena González-Castro, and Claudio et al. Carta. 2017a. Linked Registries: Connecting Rare Diseases Patient Registries through a Semantic Web Layer. 2017 (Oct. 2017), e8327980.

Pedro Sernadela, Lorena González-Castro, and José Luís Oliveira. 2017b. SCALEUS: Semantic Web Services Integration for Biomedical Applications. 41 (April 2017), 54.

Karen Smith-Yoshimura. 2018. Analysis of 2018 International Linked Data Survey for Implementers. *The Code4Lib Journal* 42 (Nov. 2018).

Rachel Thompson, Louise Johnston, and Domenica et al. Taruscio. 2014. RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research. 29 (Aug. 2014), 780–787.