A Usable Knowledge Graph Framework for Linking Health Events with Environmental Data

A Thesis

Submitted to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Albert Navarro Gallinad

ADAPT Centre for Digital Content

School of Computer Science and Statistics

Trinity College Dublin

Ireland

Supervisor: Prof. Declan O'Sullivan

Co-supervisor: Dr. Fabrizio Orlandi

June 2023

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Albert Navarro Gallinad

June 20, 2023

Acknowledgments

First and foremost I would like to thank my esteemed supervisors Prof. Declan O'Sullivan and Dr. Fabrizio Orlandi for their continuous guidance, support and patience during my PhD journey. Thanks should also go to all my colleagues in the ADAPT Centre, especially Alex Randles, for their kind help and fun moments that made my stay in Ireland an unforgettable experience. I would like to thank my colleagues from the HELICAL project for their insightful comments and suggestions. Lastly, I am also thankful to my family, partner and friends for their warm support and taking my mind off work.

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 at the ADAPT Centre for Digital Content Technology (grant number 13/RC/2106_P2) at Trinity College Dublin.

Albert Navarro Gallinad

University of Dublin, Trinity College
June 2023

A Usable Knowledge Graph Framework for Linking Health Events with Environmental Data

Albert Navarro Gallinad, Doctor of Philosophy

University of Dublin, Trinity College, 2023

Supervisor: Prof. Declan O'Sullivan
Co-supervisor: Dr. Fabrizio Orlandi

Abstract. Environmental exposures transported across air, land and water can affect our health making us more susceptible to developing a disease. Researchers studying these health-environment interactions integrate and link multiple and diverse data sources as part of their research workflows. Emerging technologies such as Knowledge Graphs (KG) can make the data integration process efficient for researchers by making the datasets interoperable. However, KG technologies are not easy to incorporate into researchers' workflows due to the technical knowledge and practical expertise required to access, explore and establish relevant links between the datasets. The major contribution of this PhD thesis is the proposed framework SERDIF (Semantic Environmental and Rare Disease data Integration Framework) that allows health data researchers themselves to directly link health data with relevant environmental data in support of their research workflows. SERDIF advances the state of the art in being the first usable KG framework, that is W3C standards-based, to be developed and implemented for the study of environmental triggers associated with rare diseases. This PhD thesis yielded two minor contributions towards improving the adoption of KG technologies and promoting transparency of research methods and data reuse towards improving the efficiency of scientific research. The first minor contribution is a step by step description of the methods and results of the evaluation approach, providing KG practitioners with a reproducible example in how to make their technologies usable for domain experts. The second minor contribution is a a collection of open source artefacts as a by-product during the development of SERDIF published to promote open science. While SERDIF has been implemented for rare disease studies, the framework has the potential to be used in other contexts to address the data integration challenges of environmental studies.

Contents

List of Tables

List of Figures

Listings

Acronyms

AA Accessibility and Availability.

AI Artificial Intelligence.

CLI Command Line Interfaces.

DEU Domain Expert Usability.

DL Deep Learning.

DPIA Data protection impact assessment.

DPO Data Protection Officer.

DQ Data Quality.

DS Data Standards.

EEA European Environment Agency.

FAIR Findable, Accessible, Interoperable and Reusable.

FHIR Fast Healthcare Interoperability Resource.

GDPR General Data Protection Regulation.

GDS Global Data Schema.

HDR Health Data Researcher.

HL7 Health Level Seven international.

IEC International Electrotechnical Commission.

ISO International Organization for Standardization.

KG Knowledge Graph.

MDI Menu-Driven Interfaces.

ML Machine Learning.

NUTS Nomenclature of territorial units for statistics.

OWL Web Ontology Language.

PROVO PROV Ontology.

QB RDF Data Cube vocabulary.

R2RML RDB to RDF Mapping Language.

RDB Relational DataBase.

RDF Resource Description Framework.

RDFS RDF Schema.

RML RDF Mapping Language.

SH Semantic Heterogeneity.

SHACL Shapes Constraint Language.

SOSA/SSN Semantic Sensor Network Ontology.

SPARQL RDF SPARQL query language.

SQL Structured Query Language.

SR StRuctural heterogeneity.

SW Semantic Web.

UCD User-Centred Design.

UI User Interface.

W3C World Wide Web Consortium.

WHO World Health Organisation.

Chapter 1

Introduction

1.1 Motivation and Scope

The United Nations General Assembly declared that access to a clean, healthy and sustainable environment to be a universal human right on the 28th of July 2022 [UN, 2022]. It has been asserted that healthier environments could have prevented almost one in every four deaths worldwide associated with environmental risk factors [WHO, 2023a,2]. Particularly, environmental exposures can make us more susceptible to diseases when inhaled, ingested or in contact with our skin [Schraufnagel et al., 2019]. Breathing the fine particulates from polluted air can cause cardiovascular [Al-Kindi et al., 2020; Rajagopalan et al., 2018], kidney [Afsar et al., 2019; Xu et al., 2021] and respiratory [Bălă et al., 2021; Kim et al., 2018] diseases, allergies [Murrison et al., 2019; Patella et al., 2018] and cancers [Institute, 2022; Turner et al., 2020]. The exposures can affect some people more than others, depending on the health context of the individual [Hooper and Kaufman, 2018; Schraufnagel et al., 2019]. For example, people with an existing health condition who are genetically predisposed could be more likely to have a relapse or health event triggered by contact with an external environmental agent. In order for researchers to propose disease prevention measures at an individual or population level, they need to integrate and link available environmental data and health information from different sources as part of their research workflows [Maitre et al., 2022; Standing Committee on Emerging Science for Environmental Health Decisions et al., 2018; Zaitchik et al., 2020]. Two key challenges arise consequently. In Fig. 1.1, the typical health-environment researcher workflow is shown at the top of the diagram, and two key challenges addressed in this thesis (and how they are approached) are situated in the "integrate available data" process of the workflow. Each of the challenges are now discussed.

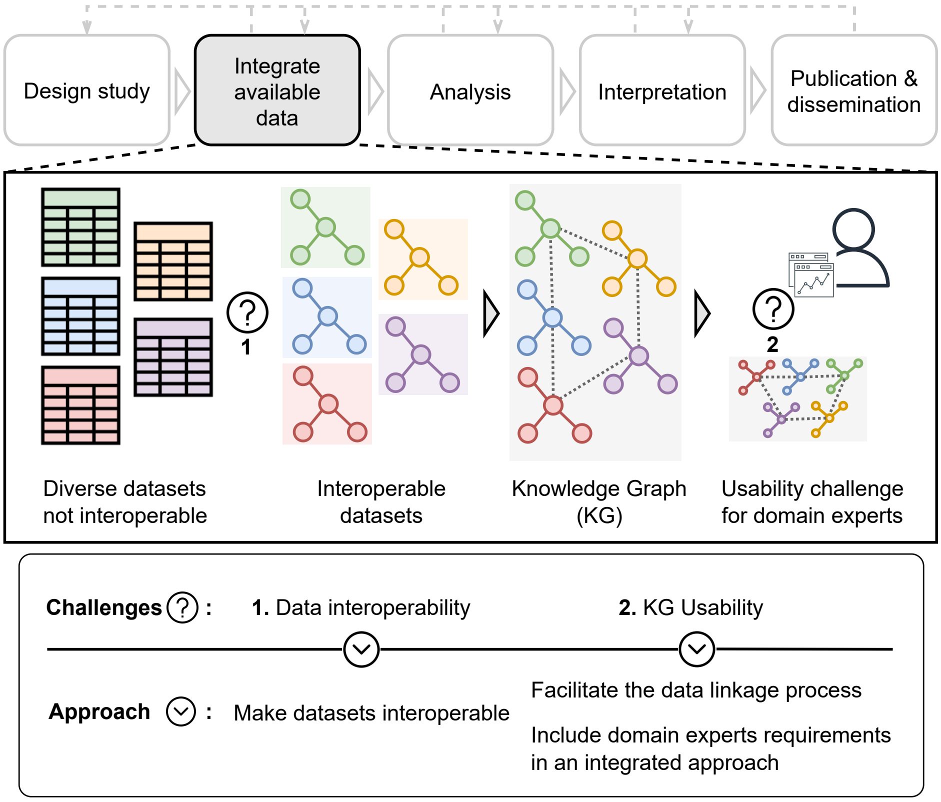

Design study
Integrate available data
Analysis
Interpretation
Publication & dissemination
1
2
Diverse datasets not interoperable
Interoperable datasets
Knowledge Graph (KG)
Usability challenge for domain experts
Challenges ? :
1. Data interoperability
2. KG Usability
Approach :
Make datasets interoperable
Facilitate the data linkage process
Include domain experts requirements in an integrated approach

Figure 1.1: Overview diagram to scope the data integration and usability challenges faced by researchers and the components of the approach to address them within an health-environmental workflow. The user icon with the plot windows represents health data researchers.

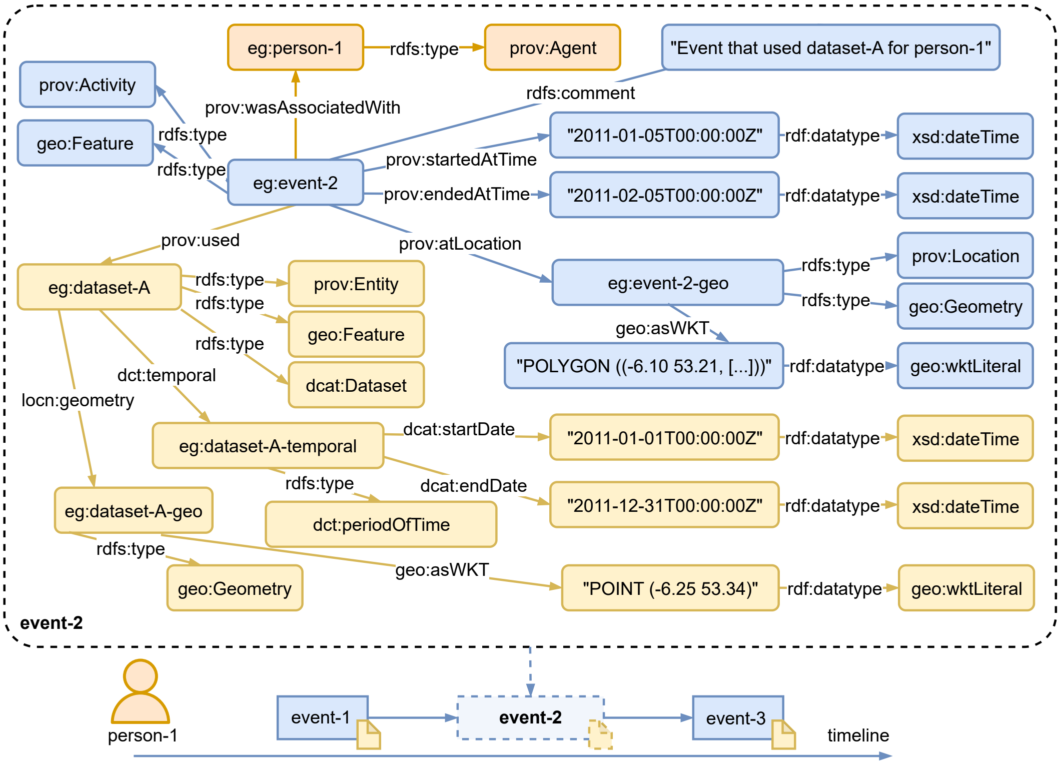
Challenge 1: Data interoperability. Researchers face significant technical challenges when integrating their complex scientific datasets [Canali and Leonelli, 2022; Ives et al., 2022; Maitre et al., 2022; Sillé et al., 2020; Standing Committee on Emerging Science for Environmental Health Decisions et al., 2018]. Potential technical challenges include the vast volume of data being generated rapidly, the integration of data from different sources, the complex data types used, the presence of duplicates, the identifier mismatches between datasets, the disconnection between data sources due to compatibility issues, the reuse of data from existing studies to validate their findings and the interdisciplinary use of data. Together these leave researchers with an overall interoperability challenge in trying to achieve an efficient data integration process.

The data interoperability challenge has been successfully addressed in other domains such as enterprise, retail, bioinformatics, biology, life sciences, public sector, etc – as Hogan, A. 2020 nicely summarised in his article [Hogan, 2020], by using a Knowledge Graph (KG) approach based on Semantic Web (SW) technologies.

A Knowledge Graph (KG) is real-world data represented using a graph-based data model, where the nodes represent entities and the edges the potential relationships between the entities [Angles et al., 2017]. In each of the success stories called out in

Hogan's article [Hogan, 2020], researchers harmonised the data that they were working with into a common graph data model. The most commonly used graph-based data models [Angles et al., 2017] include (i) the directed edge-labelled graph, where the edges represent binary relations between the entities; (ii) the heterogeneous graph, where the type of node is part of the graph model and not described as a relationship as in directed edge-labelled graphs; (iii) the property graph, where property-value pairs and labels can be associated to entities and relationships; (iv) the graph dataset, which can manage multiple graphs by using an identifier for each graph to differentiate it from the main or default graph. A KG can be stored in a graph specific data store or in relational databases (or other custom approaches), each of which aim to make the storage and retrieval of information efficient. The graph data model grants flexibility to the data integration process as additional relationships can be generated to link the data.

Semantic Web (SW) technologies refer to the World Wide Web Consortium (W3C) standards to support the Web of data with the goal of making data machine- and human-understandable [W3C, 2015]. The W3C developed the Resource Description Framework (RDF) [Cyganiak et al., 2014] as the standard graph data model, a directed edge-labelled graph, for data interchange publication of information on the Web. Thus, SW technologies are a standard-based implementation of a KG approach. RDF data is structured as a graph where real-world concepts are represented as subject – predicate – object statements (triples), capable of describing any type of data (Fig. 1.2). The entities (subject) and relationships (predicate) of the statements are identified with Uniform Resource Identifiers (URI), and the object entity is either a URI or a literal (of type string, date, integer etc.). The W3C's SW technology stack includes three languages built on RDF that promote efficient data integration and interoperability of data, RDF Schema (RDFS) [Brickley and Guha, 2014], the Web Ontology Language (OWL) [Group, 2012] and the RDF SPARQL query language (SPARQL) [W3C SPARQL, 2013]. The SPARQL specification used in this thesis is SPARQL 1.1, hereinafter referred as SPARQL. RDFS is a vocabulary description language to represent simple RDF vocabularies on the web, providing the mechanisms to describe groups of related resources (classes) and the relationships between these resources (properties). OWL can be used to describe ontologies (i.e. richer and complex vocabularies) that provide meaning and appropriate use of the data with logic rules. SPARQL can be used to express queries across diverse data sources in RDF, or viewed as RDF via a middleware, with the capabilities of querying graph patterns for tasks such as data exploration, aggregation, transformation, annotation or validation tasks to name a few. Another important feature of SPARQL is the ability to generate RDF based on query results. Researchers adopting SW technologies will promote the knowledge transfer across research groups, communities and disciplines towards advancing collaborative



```
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> . @prefix eg: <http://www.example.org> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> . @prefix prov: <http://www.w3.org/ns/prov#> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> . @prefix dbpedia:  <http://dbpedia.org/resource/> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .  @prefix locn: <http://www.w3.org/ns/locn#>.
@prefix dcat: <http://www.w3.org/ns/dcat#> .  @prefix dct: <http://purl.org/dc/terms/>
```

Figure 1.2: Example of event data linked with a dataset being described using a RDF based approach.

and open science practices [Ramachandran et al., 2021].

A standards-based KG implementation following W3C principles not only addresses the data interoperability challenge by providing meaning to the data, but it also provides querying and reasoning capabilities to generate new insights (efficiency) and transparency of the linkage process (explainability) [Hogan et al., 2022]. In order for researchers to assess if the integration has been done correctly with a KG approach, there needs to be the ability to allow researchers to pose and get answers to specific research questions. Thus it was hypothesised in this thesis's research that researchers would welcome self contained and ready to be analysed graph data as an outcome of a semantic data integration process.

Challenge 2: KG usability. While a KG approach has the potential to make a health-environment workflow efficient, it is seen in the state of the art that current usability of KG approaches is limiting the engagement and uptake by domain experts [Boyles et al., 2019; Heacock et al., 2020; Hogan, 2020; Ramírez-Andreotta et al., 2021], potentially losing the potential value for researchers trying to answer health-environment questions. Diminished usability starts when domain expert cannot access and explore the data within a KG directly themselves, which does not allow them to establish informative links between the data through their own queries resulting in subsets of the data from the KG [Rietveld and Hoekstra, 2017]. A common approach to facilitate the uptake and use of KGs by domain experts is to use visual tools and interfaces [Aguiar et al., 2021; Al-Tawil et al., 2020; Dadzie and Pietriga, 2016; Kuric et al., 2019; Pesquita et al., 2018]. It was hypothesised in this thesis's research that design of such interfaces should follow a User-Centred Design (UCD) guided by expert requirements and evaluated with usability studies [Braşoveanu et al., 2017; Desolda et al., 2020; Hogan, 2020; Wilcke et al., 2019]. In this way, usable tools and methods could enable domain experts themselves to run linkage queries on the graph data in a KG through a User Interface (UI) in an intelligible manner. Engaging domain experts with KGs would ultimately also benefit data engineers as they will be able to tailor data structures and methods to facilitate the integration step [Hogan, 2020].

In this thesis, a KG approach has been integrated within the health-environment researcher workflow to link environmental data to particular health events, guided by domain experts. The proposed approach focuses on allowing researchers themselves to select, from the available environmental datasets in a graph format, a subset that is relevant to the health events of interest to the researcher, link the relevant environmental and health data, and export the output in a form that is directly usable in the analysis stage of the researcher's workflow. In other words, raw environmental data is transformed to usable health-environmental linked data.

In this research, the environmental datasets are linked with health events by two spatiotemporal components: the distance from the health event location and a time window of data prior to the event date. Environmental exposures during time windows where individuals are more susceptible in disease development can be the most important aspect to understand health outcomes [Cui et al., 2016; Ives et al., 2022; Standing Committee on Emerging Science for Environmental Health Decisions et al., 2018]. The data linkage takes place at a query level where the location and time window are used as the common aspects to link the data for each event. A human-in-the-loop approach is needed to define the spatial and temporal aspects of the queries. A human is also needed for each use case as it improves the flexibility of the input data involved and the authoritativeness of the linkage undertaken. This contrasts with other approaches in the health-environmental domain that use automated linking, ontology matching [Euzenat and Shvaiko, 2013] or interlinking approaches in this domain, but without guaranteeing authoritativeness [Boyles et al., 2019; Ramírez-Andreotta et al., 2021].

In terms of scope, the research presented in this thesis addresses only the interoperability and usability aspects of the 'integrate available data' process within the health-environment researcher workflow (see top of Fig. 1.1). The approach also promotes collaboration between domain experts and data engineers to accomplish shared goals in this regard. The research is not focused on solving potential scalability or performance issues, nor assess the discoverability of new links beyond the researcher's requirements. Furthermore, the focus of the research in relation to the researcher's workflow is not to support processes related to the design of the study nor analyse nor interpret the data (that is the greyed-out processes of the workflow at top of Fig. 1.1). Rather, it is to provide the means for researchers to link environmental data with particular health events. The result of the KG-based linkage process is self-contained data immediately ready for analysis, with enough semantics (context) for its use in the researcher's workflow and reuse in other contexts.

1.2 Research Question

The research question examined in this thesis is:

To what extent can a Knowledge Graph (KG) framework, that is standards-based, enable Health Data Researchers (HDR) to effectively link environmental data with particular health events through location and time?

The terms in the research question are used in the following manner.

Knowledge Graph standard-based framework: combination of methods and tools based on the use of World Wide Web Consortium (W3C) standards to model graph data: the Resource Description Framework (RDF) [Cyganiak et al., 2014], the RDF query language SPARQL (SPARQL) [W3C SPARQL, 2013] and the databases to store RDF graphs, called triplestores.

Health Data Researchers: researchers that gather, analyse and link information about people and their health to enable advances in healthcare and ultimately make improvement to healthcare for all [UK, 2023]. The domain combines maths, statistics, and technology to manage and analyse very large amounts of different health data sets across our health and care systems [UK, 2023]. In this thesis the definition is narrowed to researchers studying the environmental risk factors associated with health outcomes. Health Data Researchers (HDR) are a subset of expert users (domain experts) described by Dadzie et al. 2011 [Dadzie and Rowe, 2011] as Linked Data users that *"may not necessarily have (expert) knowledge of SW technologies, but are likely to make use of sophisticated, domain specific analysis tools to manage and interact with often very large amounts of complex, heterogeneous data. They are therefore likely to have a very good understanding of data structure and content in their domain, and bring this knowledge to guide both exploratory knowledge discovery and directed information retrieval, to* *enhance their ability to obtain the insight brought to bear in decision-making."* In this thesis, the term HDR is used when describing the evaluation and implementation of the framework to particular use cases. The term HDR is generalised to domain experts (expert users) when describing the design of the framework highlighting the applicability to other domains. In the thesis, the term researcher is used as a shorthand for either HDR or domain expert depending on the context.

Enable the data linkage: data linkage is defined as bringing together from two or more different sources, data that relate to the same individual, family, place or event [Holman et al., 2008]. In this thesis enabling data linkage refers to providing the means for a Health Data Researcher (HDR) to (a) make their datasets interoperable as RDF files, (b) create queries to link and integrate datasets for a particular event based on the spatial and temporal aspects of the data and (c) export a transformed view of the linked data in a usable format for humans and machines.

Effective: producing the result that is wanted or intended; producing a successful result [Oxford, 2023].

Environmental data: any measurements or information that describe environmental processes, location, or conditions; ecological or health effects and consequences; or the performance of environmental technology [EPA, 2015]. In this thesis the focus of environmental data is on *weather and air quality observations as geolocated time series.*

Particular health events: something that occurs at a given time at a specific location [WordReference, 2023] and affects the general condition of the body or mind with reference to soundness and vigour [EUPATI, 2023]. A health event can be positive or negative and examples are the development of a disease, an injury, or responding to a medicine [Knowledge, 2023]. Health event data affecting individuals or populations includes registration of births and deaths, disease registers, routine surveys of self-reported health and health activity data from primary and secondary care [Knowledge, 2023].

1.3 Research Objectives

Research Objectives (RO) have been defined in order to support the answering of the research question of the PhD thesis (Section 1.2).

1.3.1 Research Objective 1 (RO1)

RO1: Conduct a state-of-the-art review of how to make data interoperable and usable for scientific research.

Data integration methods are reviewed towards identifying the technical challenges faced by domain experts when integrating diverse datasets in their workflows, and motivating KG and SW solutions for general and health domains (*Challenge 1: data interoperability*). As KGs tend to be machine understandable but often not human usable, common solutions to make KG accessible for domain experts are explored. In addition, special attention is given to the evaluation methods used to increase the usability of the solutions in the health domain (*Challenge 2: KG usability*).

1.3.2 Research Objective 2 (RO2)

RO2: Identify HDR requirements in linking data for health-environmental research.

The identification of the requirements started by understanding the context of the data integration and how HDRs might use the resulting linked data. A series of user requirements were specified based on the results from the state-of-the-art review supported by gathering of domain expert consensus within a case study. The requirements were then further refined and explored in each phase of the iterative evaluation process (RO4). New requirements appeared during each evaluation iterative process or previous requirements were refined, and these requirements were then incorporated in successive evaluations.

1.3.3 Research Objective 3 (RO3)

RO3: Develop a framework that enables a HDR to link environmental data with particular health events based on user data inputs.

The limitations uncovered in the state-of-the-art review (RO1) and the requirements gathered from domain experts (RO2), motivated the design of a framework called

SERDIF (Semantic Environmental and Rare Disease data Integration Framework). The framework has been developed using a User Centred Design (UCD) methodology. SERDIF is a combination of tools and processes to enable researchers to effectively link health and environmental data using a Knowledge Graph approach. The framework has three components: Knowledge Graph, Methodology and User Interface.

The **Knowledge Graph (KG)** component is the result of making the environmental datasets interoperable by uplifting them into graphs (*Challenge 1: data interoperability*), and when the raw graph data is linked and transformed to event-environmental linked data. Query templates with placeholders for user inputs are used to aggregate the relevant subset of environmental datasets (*Challenge 2: KG usability*).

The **Methodology** is a series of steps to facilitate the data integration process in health-environment workflows generating machine- and human-understandable linked data. (*Challenge 1: data interoperability and Challenge 2: KG usability*)

The **User Interface (UI)** component is designed from a user-centric perspective to support HDRs to access, explore and export of linked health-environmental data by allowing domain experts to make and run queries with appropriate visualisations. (*Challenge 2: KG usability*) The novelty and originality of SERDIF resides in how environmental data and health data inputs are linked using KGs, and the combination of the three components to generate machine- and human-understandable linked data for health-environment research.

1.3.4 Research Objective 4 (RO4)

RO4: Evaluate and refine the developed framework through rare disease case studies.

The evaluation strategy of the framework has been designed based on usability testing approaches reviewed from the state of the art (RO1) and domain expert requirements for research into the impact of the environment on health (RO2). The evaluation of the designed framework comprised a usability testing experiment with three iterative and progressive phases. The evaluation included case studies that required meaningful data linkage to test hypotheses exploring environmental risk factors, potentially providing new insights into disease aetiology.

The usability testing experiment focused on three rare disease case studies: (i) ANCA-Associated Vasculitis (AAV) in Ireland, (ii) Kawasaki Disease (KD) in Japan and (iii) AAV in Europe. AAV and KD are vasculitis that affect small and medium blood vessels, respectively, in different parts of the body in a progressive manner, resulting in damage to vital organs. While the aetiology of the vasculitis is unknown, the current theory involves a complex interaction between environmental and epigenetic factors, in a genetically susceptible individual [Kitching et al., 2020; Rodó et al., 2016,1; Scott et al., 2020]. These case studies took place in a sequential manner, with the results from each case study informing the refinement of the requirements and framework before the next case study.

1.4 Technical Approach

First, a state-of-the-art review was conducted on the existing challenges in research when integrating data from diverse sources, and the suggested solutions to address these challenges in different domains. The results of the review converged on solutions based on SW and KG approaches as candidates to effectively integrate diverse data sources in general and health related domains. However, a common challenge in applying SW and KG approaches was the difficulty in implementing them for researchers without practical expertise in the technologies (expert users [Dadzie and Rowe, 2011]). Second, a state-of-the-art review was undertaken on how to facilitate the use of KG for Health Data Researchers (HDR). In particular, on how the KG can be queried to access, explore and export relevant data based on user inputs. The review identified as a common solution the use of visual tools developed using a User Centred Design (UCD) approach, as an adequate approach for HDRs, including a usability evaluation of the solution.

As part of the UCD method, an initial set of domain expert requirements were gathered from scientific meetings and through undertaking a consensus process with HDRs in the first case study (AAV in Ireland) [Navarro-Gallinad et al., 2020]:

- *Requirement 1: Enable the HDR to query specific clinical patient data to retrieve linked environmental data, without the need for knowledge of the underpinning KG technologies;*

- *Requirement 2: Support the understanding of the HDR in the use and limitations of the linked environmental data to support identification of flares for rare diseases;*

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

The process of gathering the initial domain expert requirements is further described later in Chapter 3, and the refinement of the requirements in Chapter 5. A prototype of the KG-based approach was developed as validation of the design choices from the state-of-the-art findings and to explore the gap in the state of the art in a health use case (Section 3.3).

The initial prototype was assessed by designing a usability evaluation that combined a think aloud protocol [Boren and Ramey, 2000] while completing a series of tasks, designed with real workflows in mind, and a standard post-study questionnaire (Chapter 5). The evaluation design was also informed by the second state-of-the-art review on usability evaluations of KGs. The usability evaluation of the initial prototype found that while the visual tool provided an adequate approach to access and explore the KG, other components were needed to achieve the requirements. The state-of-the-art limitations, the initial domain expert requirements, and the evaluation results of the initial prototype provided enough information to understand the context of how Health Data Researchers (HDRs) would be using the KG-based approach within their workflows.

The understanding of the context of use arising from the initial prototyping phase, motivated the design of a framework called SERDIF (Semantic Environmental and Rare Disease data Integration Framework), which is a combination of Methodology, a KG and a UI, in support of HDRs [Navarro-Gallinad et al., 2021]. The framework was then evaluated through three usability evaluations, with a new case study for each iteration, and increasing the pool of researchers at every iteration. After the first evaluation, the interoperable data (RDF) was required as an output, in addition to the data for analysis (CSV). The RDF output was made available as increasingly the scientific community required the research outcomes to be published as Findable, Accessible, Interoperable and Reusable (FAIR) data [Wilkinson et al., 2016]. During the process of making data FAIR, the most complex step for researchers is to make data interoperable following best practices (e.g. using W3C standards) [Allen and Hartland, 2018; Belete et al., 2017; Jacobsen et al., 2020a,2]. The RDF output also includes information about the origin of the dataset together with the processing steps to generate the dataset (i.e. provenance metadata). Then, the researcher can make FAIR by: (i) licensing the dataset, specifying the data use, and (ii) defining the accessibility of the dataset and metadata when deposited in an open data repository. The interoperable data and provenance metadata were added to the framework and evaluated in the second and third evaluations.

1.5 Evaluation Approach

The usability and potential usefulness of SERDIF to link health events and environmental data for research were evaluated by conducting a usability study.

The usability study consisted of three phases (Fig. 1.3), named Phase 1 (P1), Phase 2 (P2) and Phase 3 (P3): (P1) identifying and refining the initial user requirements, (P2) validation of the usability and potential usefulness of the framework for HDRs and (P3) consolidation of the requirements and framework as a solution for HDRs. Each

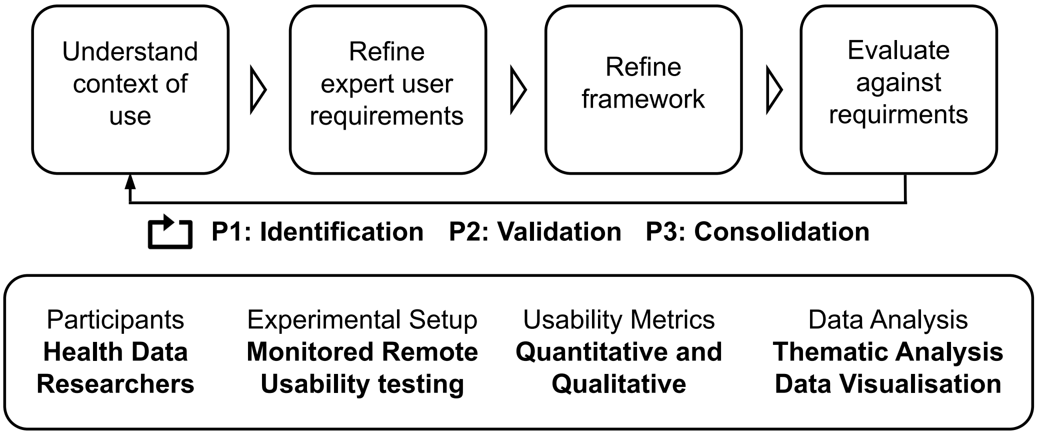


Figure 1.3: Overview of the evaluation approach following a User-Centred Design (UCD). The replay symbol represents the iterative phases P1, P2 and P3.

phase represented an iteration that comprised the development or refinement of the framework based on new requirements or refined requirements arising from the previous phase. The usability study combined summative and formative conceptualizations of usability within an iterative design process, following best practices [Lewis, 2014].

The usability tests shared static evaluation elements, which remained the same throughout the three usability tests, and dynamic evaluation elements, which were progressively incorporated in each phase of the usability study (Fig. 1.3). The static elements included the strategy (three phases), participants (HDRs), experimental setup (monitored remote usability testing), metrics (quantitative and qualitative) and data analysis (thematic analysis and data visualisation). The dynamic elements for health-environment studies of rare diseases included the description of the use case, sample size, experimental setup updates and the series of tasks to complete to link the data using SERDIF.

The usability study advanced in a progressive manner through each phase, including researchers and the use case from the previous phase. This iterative approach improved the chances to find errors, ambiguous information and confusing features while generalising the health data input capabilities of the framework. The coverage of the environmental data also increased from a single county (Ireland) in P1 and P2, to multiple countries within Europe in P3, as determined by the use cases involved.

The evaluation approach of this thesis is described in detail in Chapter 5.

1.6 Thesis Contributions

The major contribution of this PhD thesis is the proposed framework SERDIF (Semantic Environmental and Rare Disease data Integration Framework) that allows Health Data Researchers themselves to directly link health data and associated environmental data in support of their researcher workflow.

Two minor contributions of this thesis: a) the results of the evaluation approach taken to validate the usefulness of the framework for Health Data Researchers (HDR); and b) the open source artefacts (data and processes) to promote reuse, reproducibility and credibility of the framework developed.

1.6.1 Major: SERDIF

The major contribution of this thesis is the Semantic Environmental and Rare Disease data Integration Framework (SERDIF) comprising a Knowledge Graph (KG), a Methodology and a User Interface (UI). The design was informed by the limitations identified during the state-of-the-art review on how to utilise KG approach to support health-environment research, and informed by domain expert requirements from a case study. The development was guided by the iterative usability evaluation results incorporating expert feedback throughout the development process. The evaluation results are promising in that they indicate that the framework is usable and potentially useful in allowing researchers themselves to link health and environmental data whilst hiding the complexities of KG technologies.

While SERDIF has been implemented for rare disease studies, the framework has the potential to be used in other contexts to address the data integration challenges of environment related studies.

SERDIF has also been developed to comply with and promote the data governance aspect of the processing of health and environmental data, central to the linkage process. The linkage process is made transparent by providing researchers with information, in a human- and machine-understandable format, about the origin of the data and the processing steps undertaken in generation of the data. Researchers are supplied with enough context to use the data for their research.

SERDIF is already having an impact on the research community, as seen through successful publications arising out of the research.

Publications associated with this contribution are:

– **A. Navarro-Gallinad, F. Orlandi, J. Scott, M. Little and D. O'Sullivan, Evaluating the usability of a semantic environmental health data framework: approach and study. Semantic Web Journal 11(1) (2022), Publisher: IOS Press. https://doi.org/10.3233/SW-223212**

The journal article describes in detail the evaluation approach and study (second use case, see Section 1.4) of SERDIF as a guide for researchers in making KG technologies more accessible to domain experts through usability studies.

- **A. Navarro-Gallinad, F. Orlandi and D. O'Sullivan, Enhancing Rare Disease Research with Semantic Integration of Environmental and Health Data, in: The 10th International Joint Conference on Knowledge Graphs, IJCKG'21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 19–27. ISBN 978-1-4503-9565-6. https://doi.org/10.1145/3502223.3502226**

The paper describes the design of SERDIF in detail and presents the results of the first evaluation of the framework.

The paper also received a best paper nomination at the IJCKG'21 conference.

- **Albert Navarro-Gallinad, Fabrizio Orlandi, Dipak Kalra, Xavier Rodó, Mark Little and Declan O'Sullivan. Rare disease: it's all about combining data. The Marie Curie Annual Conference (MCAA). 2021. https://www.mariecuriealumni.eu/sites/default/files/2021-04/MCAA-**

  Book-of-Abstracts-2021-v7.pdf

The poster abstract presents the components of SERDIF.

- **A. Navarro-Gallinad, A. Meehan and D. O'Sullivan, The Semantic Combining for Exploration of Environmental and Disease Data Dashboard for Clinician Researchers, In Proceedings of the 5th International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 19th International Semantic Web Conference,**
  VOILA@ISWC 2020. 2778 13. http://ceur-ws.org/Vol-2778/paper7.pdf

The paper describes a preliminary user interface and an evaluation approach towards facilitating the use of KG technologies for HDRs, when linking health and environmental data. VOILA is the premier venue for HCI issues related to KGs.

1.6.2 Minor: Results of the Evaluation Approach

The first minor contribution of this thesis is the **step by step description of the methods and results of the evaluation approach**. The evaluation has been proven to be effective in improving the usability of each of the components of SERDIF (i.e.

KG, Methodology and UI). The contribution is beneficial for KG practitioners and for researchers from other domains. The detailed description and implementation of the evaluation methods provides KG practitioners a guide to improve the usability of their tools. By reducing the expertise required to benefit from KG technologies, researchers from other domains will have the opportunity to improve their own workflows. The collaboration with domain experts will then facilitate the uptake by providing tailored solutions improving the overall SW domain.

Publications associated with this contribution are shared with the major contribution of this thesis as the publications presented the framework application including the usability evaluation results.

1.6.3 Minor: Open Source Artefacts

The second minor contribution of this thesis is the collection of **open source artefacts** arising from the development of SERDIF. This is a contribution to open science promoting transparency of research methods and data reuse towards improving the efficiency of scientific research. The open source artefacts are divided into code, tools and datasets. The code related to the SERDIF UI, API and evaluation analysis have been published on GitHub and indexed in Zenodo. The tools are made available through the ADAPT centre IT services include the SERDIF UI and triplestore [Navarro-Gallinad, 2023]. The datasets deposited in open data repositories like Zenodo include an example output of the linked health-environmental data, and weather and pollution data uplifted to RDF, which also comply with the FAIR principles.

Publications associated with this contribution are:

– **Albert Navarro-Gallinad, Maria Christofidou, Solange Gonzalez Chiappe, Nathan Lea, Dipak Kalra, Fabrizio Orlandi and Declan O'Sullivan. Rare diseases: making environmental health studies' data as open as possible. The 20th International Vasculitis and ANCA Workshop. 2022. https://vasculitis2022.org/**

The poster abstract presents how researchers can make rare disease research data Findable, Accessible, Interoperable and Reusable (FAIR) whilst ensuring good data protection practices through semi-automated support.

- **Navarro-Gallinad, Albert, Orlandi, Fabrizio, and O'Sullivan, Declan. (2023). Environmental data associated to particular health events example dataset (Version 20230307) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5544257**

This published dataset includes an example output for environmental data (i.e. climate and pollution) linked with individual events through location and time (as a spreadsheet, graph and interactive report).

- **Albert Navarro-Gallinad. (2022). NUTS-RDF in GeoSPARQL (20220503T150000) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6514296**

The published dataset includes the NUTS (Nomenclature of territorial units for statistics classification) representing the regions in the EU as GeoSPARQL structures using a Construct SPARQL query (as a graph).

- **Albert Navarro-Gallinad. (2021). Weather and Air Quality data for Ireland as RDF data cube (20211221T120000) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5799118**

The published dataset includes weather and air quality data for Ireland as RDF data cube.

The following publications, although not the focus of this thesis, demonstrate the involvement in health-environmental research related to the SERDIF use cases and conducted during the PhD thesis period.

- **Scott, J., Havyarimana, E., Navarro-Gallinad, A. et al. The association between ambient UVB dose and ANCA-associated vasculitis relapse and onset. Arthritis Res Ther 24, 147 (2022). https://doi.org/10.1186/s13075-022-02834-6**

The paper analyses the effect of ultraviolet light type B (UVB) on people with ANCA-associated vasculitis towards a better understanding of the relapse events.

- **Xavier Rodó, Albert Navarro-Gallinad, Tomoko Kojima, Joan Ballester and Sílvia Borràs. Sub-weekly cycle uncovers the hidden link of aerosols and their composition to Kawasaki disease. First European Congress on Kawasaki Disease (EUROKIDS). 2021. https://www.alphavisa.com/euro-kids/2020/index.php**

This abstract presented as an oral presentation explains the time series analysis to study an epidemiological association between aerosols and Kawasaki Disease in Japan.

- **Xavier Rodó, Albert Navarro-Gallinad, Tomoko Kojima, Joan Ballester and Sílvia Borràs. Particulate matter dynamics and chemistry drive Kawasaki disease epidemiology in Japan. The 10th International Conference on Children's Health and the Environment (INCHES). 2020. https://inchesnetwork.net/conference-2020/**

The abstract presented as an oral presentation explains a study on tropospheric aerosols to understand a potential driving mechanism of Kawasaki Disease in Japan.

- **Ballester J, Borràs S, Curcoll R, Navarro-Gallinad A, Pozdniakova S, Cañas L, et al. (2019) On the interpretation of the atmospheric mechanism transporting the environmental trigger of Kawasaki Disease. PLoS ONE 14(12): e0226402. https://doi.org/10.1371/journal.pone.0226402**

The paper is a formal comment to clarify interpretation and use of a previously published approach to study the identity of the potential trigger for Kawasaki Disease.

1.7 Thesis Overview

The remainder of this thesis is structured as follows.

1.7.1 Chapter 2: Background

This chapter provides useful preliminary information for readers of this thesis. It begins with information about the use of SPARQL query templates to generate linked data. It then describes the technical complexities associated with linking health-environment data.

1.7.2 Chapter 3: State of the art

This chapter analyses the state of the art. First papers that review the state of the art related to the challenge of data interoperability are analysed. Then the state of the art with respect to user interaction with KG technologies is analysed, together with an analysis of usability studies that have been undertaken for KG-based applications in the health domain.

1.7.3 Chapter 4: SERDIF Design

This chapter describes expert user requirements that informed the design of the Semantic Environmental and Rare Disease data Integration Framework (SERDIF). This is followed by a description of SERDIF from a technology independent perspective. Finally, design considerations for the implementation by other researchers of SERDIF for other use cases are presented.

1.7.4 Chapter 5: SERDIF evaluation and implementation

This chapter describes the evaluation undertaken for the SERDIF framework over three phases. The chapter also includes the framework implementations and the expert user requirements refined at each of the three evaluation phases.

1.7.5 Chapter 6: Conclusion

This chapter presents the key findings of the research described in this thesis. It discusses to what extent the research question of this thesis has been answered and the extent to which the research objectives have been met. Possible directions for further work related to the research in this thesis are also outlined.

Chapter 2

Background

This chapter presents background information related to the research in this thesis that will aid a reader that is unfamiliar with the use of SPARQL query templates to generate linked data and the complexity associated with health-environment data. There is an assumption that the reader is familiar with the basic concepts of Knowledge Graphs (KG) and Semantic Web (SW) technologies, as discussed in Section 1.1.

This chapter starts by presenting how to build and use SPARQL queries templates to link RDF datasets following best practices (Section 2.1). This chapter follows with the description of the complexities associated with health-environment data in terms of the diverse types and scale of the data involved in data linkage tasks (Section 2.2)

2.1 SPARQL query templates

This section provides information about how to use SPARQL as a tutorial to contextualise the complexity associated with linking data with SPARQL queries. This section starts with a brief introduction on how to structure SPARQL queries and follows with two subsections describing a complex use of SPARQL query templates to link the datasets through their spatial (Section 2.1.1) and temporal features (Section 2.1.2). The example of event data displayed in Fig. 1.2 is used to facilitate and clarify the complexity associated with SPARQL queries. While only data of the event-2 is exemplified in Fig. 1.2, the queries are written with the assumption that RDF data exists for the three events presented over the timeline: event-1, event-2 and event-3.

In this thesis, SPARQL query templates are defined as follows:

SPARQL query template: a set of triple graph patterns where some of the subject – predicate – object elements include placeholders that are substituted with specific inputs.

The SPARQL language to query RDF graphs structures the queries into four sections: (1) **namespace**, (2) **query form clause**, (3) **WHERE clause** and (4) **query results modifiers**. The four query sections are exemplified in Listing 2.1 together with the results of this query on the RDF event data from Fig. 1.2. The **namespace** section describes the mapping that connects a particular prefix to a URI improving the readability of the query for humans. The **query form clause** section enables the retrieval of the results of the query as a subset of variables in the form name-value pairs (SELECT), a RDF graph constructed following a defined set of triple patterns with variables that get substituted with the query results (CONSTRUCT), a boolean value that indicates if a result exists for the query (ASK) and a RDF graph describing the resources found (DESCRIBE). The **WHERE clause** contains the triple patterns to match against the available RDF data. The **query results modifiers** can order, specify a selection, remove duplicates, limit, filter by value or be grouped by a specific variable(s) to aggregate the results from a SPARQL query.

Listing 2.1: Example CONSTRUCT query form to link a particular dataset with events.

```
# -- 1. Namespaces ------------------------------------
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX eg: <http://www.example.org/>
# -- 2. Query form clause ------------------------------
CONSTRUCT {
    ?event rdfs:type prov:Activity ;
          prov:used eg:dataset-A .
    eg:dataset-A rdfs:type prov:Entity .
}
# -- 3. WHERE clause -----------------------------------
WHERE {
    ?event prov:wasAssociatedWith ?person .
    ?person rdfs:type prov:Agent
}
# -- Query results as RDF triples ----------------------
|   subject     | predicate |   object     |
| ------------ | --------- | ------------- |
| eg:event-3    | prov:used | eg:dataset-A  |
| eg:event-3    | rdfs:type | prov:Activity |
| eg:event-2    | prov:used | eg:dataset-A  |
| eg:event-2    | rdfs:type | prov:Activity |
| eg:event-1    | prov:used | eg:dataset-A  |
| eg:event-1   | rdfs:type | prov:Activity |
| eg:dataset-A | rdfs:type | prov:Entity   |
```

From the query forms available in the SPARQL language, in this section the interest is targeted to the **CONSTRUCT query forms**, which can generate new links based on a defined set of triple patterns. For example, a particular dataset can be linked with a group of events as in Listing 2.1. This example presents the case of a custom resource (*eg:dataset-A*) being linked arbitrarily with all the events available. This resource is added manually in the CONSTRUCT clause. This capability of manually adding a custom resource is valid as long as it is written in a valid RDF syntax and does not have to be an existing resource in the dataset that is being queried. While this capability can be useful in certain situations, it is prone to errors due to the manual input component. It is recommended to validate the resulting RDF graph against the rules of the vocabularies and ontologies being used, in this case with PROV-O [Lebo et al., 2013], DCAT [Albertoni et al., 2020] and GeoSPARQL [Perry et al., 2012].

The following two subsections describe how to include new information and links generated through spatial and temporal reasoning following standards.

2.1.1 Spatial Linkage

The Open Geospatial Consortium (OGC) is an organisation with the goal of making location information Findable, Accessible, Interoperable and Reusable (FAIR) that represents more than 500 businesses, government agencies, research organisations and universities. OGC defined a standard vocabulary for representing and querying geospatial data in RDF called GeoSPARQL [Perry et al., 2012]. GeoSPARQL defines the class *geo:Feature* to represent a unique identifiable phenomenon that is bounded (e.g. a hospital or a tree). The boundaries from a *geo:Feature* can be precisely defined (e.g. an administrative region), vague (e.g. a river) or evolve in time (e.g. a person location), as described by the OGC. Then, a specific geometry (*geo:Geometry*) can be associated with the geo:Feature with relationships such as *geo:hasGeometry*, *prov:Location* or *locn:geomety* (Fig. 1.2).

Once the RDF graphs are described following OGC standards (Fig. 1.2), GeoSPARQL functions can be used to link geometries based on geographical relationships. Examples include selecting geometries that lay within, intersect, touch or contain other geometries among others. Listing 2.2 used a GeoSPARQL function to filter the datasets within the geometry of the event, and then a CONSTRUCT query form to establish a new link based on spatial relationships (i.e. spatial reasoning).

Another example of how CONSTRUCT queries and GeoSPARQL can be combined to generate new RDF graphs is one of the open source artefacts generated in this thesis (Section 1.6.3). Following the OGC standards, the NUTS (Nomenclature of territorial units for statistics classification) representing the regions in the EU [eurostat, 2021] have been converted from NeoGeo [Martín Salas and Harth, 2012] to GeoSPARQL structures using a Construct SPARQL query. The purpose of the conversion was to enable GeoSPARQL spatial reasoning features for geometry-based queries such as the one in Listing 2.2 (*geof:sfWithin*).

Listing 2.2: Example a CONSTRUCT query form to spatially link events and datasets.

```
# -- 1. Namespaces ------------------------------------
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX locn: <http://www.w3.org/ns/locn#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX eg: <http://www.example.org/>
# -- 2. Query form clause ------------------------------
CONSTRUCT { ?event prov:used ?dataset . }
# -- 3. WHERE clause -----------------------------------
WHERE {
# -- 3.1 event geometry --------------------------------
    ?event a prov:Activity, geo:Feature ;
           prov:atLocation ?eventLoc .
    ?eventLoc a geo:Geometry ;
           geo:asWKT ?eventGeo .
# -- 3.2 dataset geometry ------------------------------
    ?dataset a dcat:Dataset, geo:Feature ;
       locn:geometry ?datasetLoc .
    ?datasetLoc a geo:Geometry ;
       geo:asWKT ?datasetGeo .
# -- 3.3. Filter datasets within the geometry of the event --
    FILTER(geof:sfWithin(?datasetGeo, ?eventGeo))
}
# -- Query results as RDF triples ----------------------
|   subject     | predicate |   object      |
| ------------ | --------- | ------------- |
| eg:event-2   | prov:used | eg:dataset-A  |
```

2.1.2 Temporal Linkage

SPARQL includes functions that operate on XML schema data types [Biron and Malhotra, 2004], W3C recommendation for data types, including integers, decimals, floats, doubles, strings, booleans and date time values. Examples of these functions include expressions to subset, replace, concatenate or round the literal values of a particular data type. Functions on date and times grant the possibility to select each of the components that are part of a date time literal value (*xsd:dateTime*). In addition, logical expressions can be described in SPARQL FILTERS to restrict the solutions of a graph pattern to specific constraints. Listing 2.3 restricts the temporal linkage of events with datasets based on an overlapping interval between the two resources. The complexity of the constraints depends on the type of temporal linkage needed, which at the same time can be combined with spatial constraints or other types of constraints.

Beyond *xsd:dateTime* in SPARQL queries, another option is to describe the temporal aspects of RDF with OWL-Time [Cox and Little, 2022]. OWL-Time ontology describes the temporal properties of resources with a vocabulary to order time resources based on time duration and positions. The benefits of using this ontology include the description of the data and time with the resolution that is known, which can be very useful in the health domain. For example, if a person is not sure when the symptoms started, the information can be recorded as an event that happened in a particular month without including further unknown information such as the day and time. In this manner, the human or machine exploring the data would be able to use and interpret the event data in an appropriate manner. However, OWL-Time is a draft of a potential specification (i.e. future W3C recommendation) as of the submission of this thesis.

Listing 2.3: Example a CONSTRUCT query form to temporally link events and datasets.

```
# -- 1. Namespaces ------------------------------------
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX eg: <http://www.example.org/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#> .
# -- 2. Query form clause -----------------------------
CONSTRUCT { ?event prov:used ?dataset . }
# -- 3. WHERE clause ----------------------------------
WHERE {
# -- 3.1 event time interval --------------------------
   ?event a prov:Activity ;
      prov:startedAtTime ?evStTime ;
      prov:endedAtTime ?evEndTime .
# -- 3.2 dataset time interval ------------------------
   ?dataset a dcat:Dataset ;
```

```
       dct:temporal ?datasetTemp .
    ?datasetTemp a dct:periodOfTime ;
       dcat:startDate ?dsStTime ;
       dcat:endDate ?dsEndTime .
# -- 3.3. Filter datasets with data in the event interval --
    FILTER(?dsStTime <= ?evStTime && ?dsEndTime >= ?evEndTime)
}
# -- Query results as RDF triples ----------------------
|   subject    | predicate |   object      |
| ------------ | --------- | ------------- |
| eg:event-2   | prov:used | eg:dataset-A  |
```

2.2 Complexity of health-environment data

This section describes the complexity associated with using health-environment data for domain experts.

Data complexity refers to information that has been composed from multiple, large and diverse data sources, often referred to as "Big Data" [Andreu-Perez et al., 2015]. These data sources can vary in terms of structure, size, query language and data type, making it difficult for domain experts to efficiently integrate the data and transform it into meaningful insights [Pastorino et al., 2019]. The Health Directorate of the Directorate-General for Research and Innovation at the European Commission (EC) organised a workshop to define an action plan for Big Data in health research in 2015 that included stakeholders with diverse health domain expertise [Auffray et al., 2016]. The stakeholders reached a consensus for a workable definition for the meaning of Big Data in the health domain:

"Big data in health encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points."

Datasets complexity. The complexity of the health-environment data used in this thesis involved health data from disease registries, national surveys and administrative data, and environmental data from European environmental agencies. The data types were available in different formats and structures as the healthcare industry, government and environmental agencies target distinct goals and audiences. The initial datasets of this thesis included spreadsheets (CSV), text files (TXT), websites (HTML), compressed formats (NetCDF), relational databases (RDBMS) and graph databases (triplestore, RDF). The size of these files ranged from a few kB for the text and websites to MB and GB for the spreadsheets, databases and compressed files formats. This variety of data formats and file sizes requires the use of different query languages or data mining methods to retrieve relevant information for research.

The structured files that organised the data in a tabular format or that were converted to temporary tables during the processing step such as the compressed files did not often have a direct relationship between the columns and rows. The column names for the variables in the health and environmental data did not match and thus, the datasets could not be aligned smoothly. The temporal and spatial dimensions were shared across the datasets as date and geographical coordinates columns. However, the resolution of time and location varied across datasets including hourly, daily, monthly and yearly data (time); and regularly (i.e. gridded data) or irregularly spaced data points (location). The interpretation of the environmental measurements had to account for the differences in units, measurement equipment and meaning of the measure in a given context. This contextual information was available in the form metadata stored as tables, text or website formats.

The datasets initially used in this thesis are summarised in Table 2.1 in terms of the complexity levels described in this section. The size and variability values in this table represent the values of the health related data sources (i.e. disease registry and national survey data) computed in January 2022. The health data sources required a preprocessing step to extract particular health events relevant for a study, which was conducted by the Health Data Researchers (HDR). The environmental data presented in Table 2.1. comprises only a period of data within 2011-2021. In addition, each of the data sources in this table was accompanied by a set of metadata information which is not included in the table.

Datasets linkage points. Health and environmental data are complex on their own domain but the complexity can be even higher for domain experts when linking these diverse types of data. The common elements between health and environmental observations are time and location as the rest of the dimensions and measures do not match. As previously mentioned, there is no direct relationship between these elements unless the temporal resolution is the same and the datasets are collected in the same location. New relationships need to be established between the datasets informed by the domain experts to be adequate for each use case. For example, datasets can be linked by using the closest data points in terms of location or aggregated within an area of study at a particular temporal resolution relevant to the health outcome.

Table 2.1: Summary of the initial data sources to highlight the complexity of the data used in this thesis.

Data type [source]	**Format**	**Size**	**Variability**	**Structure**	**Temporal resolution**	**Spatial coverage**
Weather [Copernicus, 2020]	NetCDF HTML TXT	1GB	9 variables	Spatial grid time series	daily	Europe
Air pollution [EEA, 2022]	CSV	32GB	462 variables	Single location time series	daily, monthly, yearly	Ireland, UK, Switzerland, Czech Republic
Administrative areas [Navarro-Gallinad, 2022]	RDF	2MB	1782 geometries	Nested geometries	-	Europe
Disease registry [AVERT, 2022]	RDBMS RDF	16MB	668 patients	Patient records	daily	Ireland
National survey [Makino et al., 2019]	CSV	2MB	48 regions	Single location time series	daily	Japan
Disease registry [FAIRVASC, 2022]	RDBMS RDF	33MB	1391 patients	Patient records	daily	Ireland, UK, Switzerland, Czech Republic

Chapter 3

State of the Art

This chapter presents a state-of-the-art review related to the two key challenges identified for this thesis: how to make data interoperable and usable for scientific research.

The chapter starts by defining the concept of data interoperability and the aspects of the challenges associated with the process of making data interoperable (Section 3.1). Then an analysis (Section 3.1.1) according to the data interoperability challenge aspects is presented for a number of review papers that cover the state of the art for general and diverse domains. In addition, proposed solutions from the review papers to address the challenge aspects are presented. Section 3.1.2 presents a similar analysis for papers that review the state of the art in the specific health domain, of focus in this thesis. Section 3.1.3 then discusses the key findings arising from the analyses presented in Section 3.1.1 and 3.1.2 that informed the design decisions for the framework solution that the author of this thesis developed and evaluated.

The chapter then presents the state-of-the-art approaches for making Knowledge Graphs (KGs) usable for domain experts (Section 3.2). Visual tools such as query builders are first introduced as general state-of-the-art solutions to interact with the complex data in a KG (Section 3.2.1), followed by an overview of user interactions tools for rare disease and environmental research (Section 3.2.2) and then, empirical usability evaluations are reviewed for researchers in the health domain (Section 3.2.3). The findings from these two state-of-the-art reviews which informed the design of the UI implementation and evaluation approach adopted by the author of this thesis for the research, are then presented in Section 3.2.4.

In addition, the findings from the analyses of the state of the art with respect to the two challenges focused upon in this research, *Challenge 1: data interoperability* and *Challenge 2: KG usability*, have been explored in a preliminary study that evaluated the usability of a prototype KG-based approach using W3C standards for a health-environment linkage case study. The study and findings are presented in Section 3.3.

The chapter concludes with the summary of the findings from the state-of-the-art review, together with the results of the findings from the preliminary usability study of the prototype KG-based approach (Section 3.4).

3.1 Challenge 1: data interoperability

This section presents a state-of-the-art review of the existing approaches to make data interoperable to address the technical challenges faced by researchers when integrating diverse datasets as part of their workflows. *Interoperability* is defined as follows by the International Organization for Standardization/ International Electrotechnical Commission (ISO/IEC):

Interoperability: *"Ability of two or more systems or applications to exchange information and to mutually use the information that has been exchanged"* – ISO/IEC 17788:2014
Data interoperability: *"interoperability concerning the creation, meaning, computation, use, transfer, and exchange of data"* – ISO/IEC 20944-1:2013

This definition of *interoperability* has also been endorsed by the Healthcare Information and Management Systems Society in the healthcare domain [HIMMS, 2020], which further categorised *interoperability* in four levels:

1. Foundational (Level 1). Ability of a system to securely send and receive data from another system.

2. Structural (Level 2). Ability of a system to interpret the data exchange at the data field level (i.e. format, syntax and organisation).

3. Semantic (Level 3). Ability of a system to provide a shared understanding and meaning of information to another system or user by using common underlying models and codification of the data.

4. Organisational (Level 4). Governance, policy, social, legal and organisational considerations to facilitate the secure, seamless and timely communication and use of data both within and between organisations, entities and individuals. These components enable shared consent, trust and integrated end-user processes and workflows.

The ISO/IEC also defined the syntactic (i.e. structural – level 2) and semantic interoperability in the following standard document for information technology:

Syntactic interoperability: *"interoperability such that the formats of the exchanged information can be understood by the participating systems"* – ISO/IEC 19941:2017
Semantic interoperability: *"interoperability so that the meaning of the data model within the context of a subject area is understood by the participating systems"* – ISO/IEC 19941:2017

A researcher can make their relevant datasets interoperable at a basic level (i.e. levels 1 and 2) or higher level (i.e. level 3 and 4) based on the goals of their research project. Higher levels of data interoperability promote collaborative and open research practices towards transitioning from more traditional and individualistic methods of research [Oldman and Tanase, 2018]. However, making data interoperable is a complex process for domain experts, getting progressively more complex in higher levels of interoperability [Allen and Hartland, 2018; Belete et al., 2017; de Mello et al., 2022; Jacobsen et al., 2020b]. The existing challenging aspects in conducting data integration processes have been identified for the general (Section 3.1.1) and health (Section 3.1.2) domains from conducting a state-of-the-art review.

The research for review papers included scientific literature in Google Scholar considered as *review papers* with the following keywords *data integration, interoperability* and *domain expert/researcher*. Each state-of-the-art *review paper* studied recent progress regarding the data interoperability and integration aspects in a particular domain based on already published research from 2015 to 2022.

The main state-of-the-art aspects identified in the reviewing process are summarised as follows.

Domain Expert Usability (DEU). DEU refers to the technical barrier when using a system in order to achieve specific goals for research in a given context [Jokela et al., 2003]. This challenge aspect was attributed to review papers that explicitly mention the usability challenge when using a particular approach or technology to integrate heterogeneous datasets. This challenge aspect also covered references to the minimum adoption of the technology in a research domain, the need for usability evaluations to improve the uptake, the training needed to a group of researchers to be able to use a tool, as well as the prerequisite for domain expert collaboration with knowledge engineers to develop a usable solution.

Semantic Heterogeneity (SH). SH involves the inconsistency in the meaning, interpretation and intended use of a combination of data values that stemmed from diverse databases or datasets developed by independent parties [de Mello et al., 2022].

This challenge aspect was associated with review papers that referred to a misalignment between schemas (e.g. different constraints applied to data), ontologies (e.g. missing correspondences between ontologies) and data values (e.g. multiple names for the same entity) in the data integration process.

StRuctural heterogeneity (SR). SR describes the obstacle faced by researchers when datasets come in different format, syntax and organisation towards an efficient data integration process [Zaveri et al., 2016]. This challenge aspect was associated with review papers that referred for example to problems of integration of structured and unstructured datasets (format), data observations in different units, languages or date formats (syntax); and datasets generated by different entities (organisational).

Data Quality (DQ). DQ refers to the situation when data does not achieve the researchers' requirements for an intended purpose [Zaveri et al., 2016]. DQ has six main dimensions: (i) accuracy (verifiable source), (ii) completeness (include necessary data values), (iii) consistency (same data values for the same observation across sources), (iv) validity (within an appropriate range of values for a given variable), (v) uniqueness (no duplicates) and (vi) timeliness (readily available and accessible). This challenge aspect was associated with review papers that referred to any of these DQ concerns.

Accessibility and Availability (AA). AA describes the inability of researchers to find and use the relevant data for their workflows [Kamdar et al., 2019]. This challenge aspect was associated with review papers that referred to lack of metadata describing the actual data, the difficulty in using an unfamiliar dataset because of the content and/or format, and for the lack of readily available data published by independent entities in an interoperable format.

Less frequent challenge aspects that presented once or twice across the review papers but that were not included in the analysis of the state-of-the-art include: large data volume requirements for the Machine Learning (ML) related papers [Gligorijević and Pržulj, 2015; Grapov et al., 2018], lack of published research [Drury et al., 2019], how to automatise systems for research [Chakraborty et al., 2017; Krishnakumar, 2002] or data privacy [Heacock et al., 2020].

The analysed state-of-the-art review papers also suggested solutions to adopt for future research in order to address the main existing problems. The suggested solutions are summarised as follows.

Semantic Web (SW) technologies. SW technologies refer to the W3C standards to support the Web of data with the goal of making data machine- and human-understandable [W3C, 2015]. The SW technologies have been presented in Section 1.1 of this thesis.

Machine Learning (ML) and Deep Learning (DL). ML is a field in Artificial Intelligence (AI) that focuses on the use of data to build algorithms that can learn patterns from the data in a gradual manner. DL is a subgroup of ML algorithms that use artificial neural networks with a logical structure similar to the human brain. Both methods include the use of statistical methods to train algorithms to predict and classify data based on the learned patterns or rules. The learned patterns can link heterogeneous data sources by aligning data fields, schemas and ontologies; and to extract data from unstructured data sources making it available to store in a structured format to further link it with other relevant data sources.

Intelligent Systems (IS). Intelligent systems emulate some aspects of intelligence present in nature to solve complex problems more efficiently [Krishnakumar, 2002]. These aspects include learning, adaptability, robustness across problem domains, improving efficiency (over time and/or space), information compression (data to knowledge) and extrapolated reasoning.

Knowledge Graphs (KG). A KG is real-world data represented using a graph-based data model, where the nodes represent entities and the edges the potential relationships between the entities [Hogan et al., 2022]. The KGs have been presented in Section 1.1 of this thesis.

Data Standards (DS) and Global Data Schema (GDS). DS and GDS have the goal of integrating the data by normalising and standardising the datasets at a given point of the researchers workflows [Golriz Khatami et al., 2020]. Datasets can be harmonised to conform to a schema and format at the start, when data is generated or collected, or at later stages as analysis and publication. Once the data is represented following a standard or global data schema, the data is ready to be linked with other datasets adopting the same standard, even if the data is generated by independent entities.

Fast Healthcare Interoperability Resource (FHIR). FHIR is Health Level Seven international (HL7) standard for exchanging healthcare information between computer systems towards advancing data interoperability [HL7, 2022]. FHIR is founded on Web standards and uses XML and JSON formats to represent the data following the HL7 (v2 and v3) previous standards. FHIR datasets generated from independent entities can be integrated between them since they have been harmonised to a common standard for data exchange.

The data interoperability challenge aspects and suggested solutions from the review papers are organised into two subsections. The first subsection (Section 3.1.1) analyses two state-of-the-art review papers that are domain-independent and review papers addressing the challenge in four diverse domains. The second subsection (Section 3.1.2) focuses on state-of-the-art review papers on data interoperability challenge aspects in the health domain only. Section 3.1.3 then discusses the key findings arising from the analyses presented

3.1.1 General and diverse domains

This section included six *review papers* that reviewed or surveyed scientific literature on addressing the data interoperability challenge. The *review papers* are summarised in Table 3.1 indicating the type of review and topic of the paper, the total number of reviewed work analysed in the paper and the years of the reviewed work. The authors for this thesis tried to be accurate in the review work number and years but gathering the exact information for some of the *review papers* was challenging as it was not explicitly stated. In those cases, the authors based the results in summary tables or manually counting the citations referring to work being reviewed instead of cited to provide evidence to support the assertions and claims. The types of reviews included in this thesis are defined following the definitions from Grant and Booth et al. 2009 [Grant and Booth, 2009]:

Literature review. *"Generic term: published materials that provide examination of recent or current literature. Can cover a wide range of subjects at various levels of completeness and comprehensiveness. May include research findings."*

Overview. *"Generic term: summary of the [medical] literature that attempts to survey the literature and describe its characteristics."* The term *survey paper* is considered in this thesis as an *overview paper* since it was the preferred term by the authors of the reviews.

Systematic review. *"Seeks to systematically search for, appraise and synthesis research evidence, often adhering to guidelines on the conduct of a review."*

Scoping review. *"Preliminary assessment of potential size and scope of available research literature. Aims to identify nature and extent of research".*

The *review papers* also included comparative studies, defined by Coccia and Benati et al. 2018 [Coccia and Benati, 2018], and research agendas, defined by the Cambridge dictionary [Cambridge, 2023]:

Comparative study. *"Investigations to analyse and evaluate, with quantitative and qualitative methods, a phenomenon and/or facts among different areas, subjects, and/or objects to detect similarities and/or differences."*

Research agenda. *"A list of matters to be discussed on a detailed study of a subject, especially in order to discover (new) information or reach a (new) understanding".*

The state-of-the-art *review papers* were analysed according to the aspects for the data interoperability challenge previously presented and summarised in Table 3.2. The analysis included the six *review papers* from Table 3.1 that described the challenge across one general domain-independent domains (Extract, Transform, Load (ETL) [Chakraborty et al., 2017]) and four diverse specific domains (agriculture [Drury et al., 2019; Evans et al., 2017]; geoscience [Gil et al., 2018], Internet of Things (IoT) and industry [Kamm et al., 2021]; and smart grids [Wang et al., 2021]). The variety of the domains reinforce the point that making data interoperable for research use is a pressing challenge across the scientific community.

Table 3.1: Summary of the review papers that describe the data interoperability challenge in general and diverse domains.

Publication	**Type of review and topic**	**Reviewed work (#)**	**Years**
[Chakraborty et al., 2017]	A survey paper that presents a holistic view of literature in data integration and Extract-Transform-Load techniques	29	2009-2016
[Evans et al., 2017]	A literature review on distributed information technologies describing the crop-environment-management interaction	112	1975-2017
[Gil et al., 2018]	A research agenda on intelligent systems that will result in fundamental new capabilities for understanding the Earth system.	12	2012-201
[Drury et al., 2019]	A survey paper on the application of semantic web technology and data interchange protocols for agricultural problems	76	2002-2018
[Kamm et al., 2021]	A literature review on knowledge discovery in heterogeneous and unstructured data of Industry 4.0 systems	37	1996-2020
[Wang et al., 2021]	A survey paper on the development status and application prospects of knowledge graphs in smart grids	56	2005-2020

Most of the state-of-the-art *review papers* (5 out of 6, Table 3.2) reflect the difficulty for domain experts to use and combine the data for domain expert purposes (DEU) and how semantic disparities appear when working in collaborative environments that integrate data generated at different sources (SH). A few of the review papers (2 or 3 out of 6, Table 3.2) state the obstacles when finding data for the intended purpose (DQ), without the proper format and syntax to be integrated into the researchers workflows (SR) and data ready to be accessed and used for analysis (AA).

The state-of-the-art *review papers* suggested AI approaches to address these existing data interoperability challenges when integrating heterogeneous data for research. The suggested approaches from the reviews include general Semantic Web (SW), Data Standards (DS), Intelligent Systems (IS) and Knowledge Graph (KG) approaches. However,

Table 3.2: Summary analysis of the art review papers of several diverse domains according to the data interoperability challenge aspects and suggested solutions. The challenge aspects include Domain Expert Usability (DEU), Semantic Heterogeneity (SH), Data Quality (DQ), StRuctural Heterogeneity (SR), Accessibility and Availability (AA). The suggested solutions include Semantic Web (SW), Data Standards (DS), Intelligent Systems (IS) and Knowledge Graphs (KG). The black circle indicates the presence of the existing challenge aspect and the blank circle the absence of it. The KG and SW suggested approaches are bolded to highlight the design choices of this thesis.

Publication	**Domain**	**Existing challenge aspects**					**Suggested**
		DEU	**SH**	**DQ**	**SR**	**AA**	**approach**
[Chakraborty et al., 2017]	ETL	●	●	●	○	●	**SW**
[Evans et al., 2017]	Agriculture	●	○	○	○	○	DS
[Gil et al., 2018]	Geoscience	●	●	●	●	○	IS
[Drury et al., 2019]	Agriculture	●	●	○	○	○	**SW**
[Kamm et al., 2021]	IoT and Industry	●	●	○	●	●	**SW**
[Wang et al., 2021]	Smart grids	○	●	●	●	○	**KG, SW**

most of the reviews (4 out of 6, Table 3.2) suggested a SW approach after analysing recent research in the respective domain.

3.1.2 Health domain

This section included eleven *review papers* that reviewed or surveyed scientific literature on addressing the data interoperability challenge in the health domain. The *review papers* are summarised in Table 3.3 indicating the type of review and topic of the paper, the total number of references cited in the paper and the years of scientific literature included as references. The review papers terminology defined in Section 3.1.1 is also used in Table 3.3 for the categorisation of the reviewed research.

The analysis of state-of-the-art review papers according to the aspects for the data interoperability challenge in the health domain is presented in Table 3.4. This table follows the same structure as Table 3.2 to facilitate comparison and follow up discussion. The state of the review papers covered the following diverse subdomains in the health domain due to the availability of this literature in the last seven years: biological data [Gligorijević and Pržulj, 2015], precision medicine [Grapov et al., 2018], medical informatics [Ulrich et al., 2022], healthcare [Abedjan et al., 2019; Dhayne et al., 2019], biomedicine [Kamdar et al., 2019], rare diseases [Schaaf et al., 2020], biomedical data science [Callahan et al., 2020], health-environmental [Heacock et al., 2020], Alzheimer's disease [Golriz Khatami et al., 2020], and health records standards [de Mello et al., 2022].

Table 3.3: Summary of the review papers that describe the data interoperability challenge in the health domain.

Publication	Type of review and topic	Reviewed work (#)	Years
[Gligorijević and Pržulj, 2015]	A survey paper on recent methods for collective mining (integration) of various types of networked biological data.	45	2000-2015
[Grapov et al., 2018]	A literature review on the challenges and opportunities for DL at a systems and biological scale in precision medicine readership	58	2001-2018
[Abedjan et al., 2019]	A book chapter that overviews the needs, opportunities, recommendations and challenges of using (Big) Data Science technologies in the healthcare sector	63	2001-2019
[Dhayne et al., 2019]	A comprehensive survey in search of Big Medical Data integration solutions	120	2001-2019
[Kamdar et al., 2019]	A overview on the opportunities of using Semantic Web technologies and Life Sciences Linked Open Data to integrate biomedical data and knowledge from heterogeneous data sources	88	1993-2019
[Schaaf et al., 2020]	A scoping review of clinical decision support systems for rare disease patients.	22	2008-2018
[Callahan et al., 2020]	A survey on systems that formally represent knowledge to address data science problems in clinical and biological domains, and approaches of creating Knowledge Graphs.	83	2018-2019
[Heacock et al., 2020]	A literature review on opportunities to leverage data and efforts to advance data sharing and reuse across health-environmental research	65	2007-2019
[Golriz Khatami et al., 2020]	A literature review on challenges of integrative disease modelling in Alzheimer's disease	20	2010-2019
[Ulrich et al., 2022]	A systematic review on understand the definition of metadata and the challenges resulting from metadata reuse in medical informatics	81	2008-2018
[de Mello et al., 2022]	A systematic literature review on semantic interoperability in health records standards	28	2010-2020

Table 3.4: Summary analysis of the art review papers of the health domain according to the data interoperability challenge aspects and suggested solutions. The challenge aspects include Domain Expert Usability (DEU), Semantic Heterogeneity (SH), Data Quality (DQ), StRuctural Heterogeneity (SR), Accessibility and Availability (AA). The suggested solutions include Semantic Web (SW), Machine Learning (ML), Deep Learning (DL), Artificial Intelligence (AI), Knowledge Graphs (KG), FHIR and Global Data Schema (GDS). The black circle indicates the presence of the existing challenge aspect and the blank circle the absence of it. The KG and SW suggested approaches are bolded to highlight the design choices of this thesis.

Publication	**Domain**	**Existing challenge aspects**					**Suggested approach**
		DEU	**SH**	**DQ**	**SR**	**AA**	
[Gligorijević and Pržulj, 2015]	Biological data	○	○	●	●	●	ML
[Grapov et al., 2018]	Precision medicine	●	●	○	○	○	DL
[Abedjan et al., 2019]	Healthcare	●	○	●	●	●	ML, SW
[Dhayne et al., 2019]	Healthcare	●	●	●	●	○	ML, SW
[Kamdar et al., 2019]	Biomedicine	●	●	○	○	●	**SW**
[Schaaf et al., 2020]	Rare diseases	●	○	○	○	○	FHIR
[Callahan et al., 2020]	Biomedical data science	●	●	●	○	●	**KG, ML**
[Heacock et al., 2020]	Health-environmental	●	○	○	●	●	**SW**
[Golriz Khatami et al., 2020]	Alzheimer's Disease	○	●	●	●	○	GDS
[Ulrich et al., 2022]	Medical informatics	●	●	●	●	○	**SW**
[de Mello et al., 2022]	Health records standards	●	●	○	●	●	**SW**

Most of the state-of-the-art review papers for each of the topic areas refer to usability problems for domain experts DEU (9 out of 11, Table 3.4) and semantic heterogeneity SH (8 out of 11, Table 3.3) as one of the main problematic challenge aspects when integrating heterogeneous data sources. The rest of the challenge aspects are also present in half of the reviews (6 or 7 out of 11, Table 3.4), providing further indications that they are common aspects of the data interoperability challenge in the health domain.

In this domain, the suggested approach by the state-of-the-art review papers to address these challenge aspects in future studies, leans towards SW technologies, being present in half of the reviews (6 out of 11, Table 3.4). Other AI-based approaches such as ML (4 out of 11, Table 3.4), KG (1 out of 11, Table 3.4) and DL (1 out of 11, Table 3.4) are also suggested as solutions. Additionally, standard based solutions such as Fast Health Interoperability Resources (FHIR) [HL7, 2022] and a Global Data Schema (GDS) are proposed for rare disease and Alzheimer's research.

3.1.3 Discussion

From the analysis of the promising solutions put forward by the various state-of-the-art review papers discussed in Section 3.1.1 and 3.1.2, it is clear that SW technologies, which are a W3C standard approach of implementing KG, should be a preferred solution to address *Challenge 1: Data Interoperability*. The standards based approach underpinning SW technologies provides the basis for Knowledge Graphs to scale, and also make the data that is needed for interdisciplinary research interoperable. It should be noted that the selected state-of-the-art review papers included KG and SW specific reviews as they were identified from the state of the art keywords and filters used in the search, which could have introduced a small bias towards these solutions.

This solution is preferred in both general/diverse domains (4 out of 6, Table 3.2) and health domain (6 out of 11, Table 3.4), with a total of 10 out 17 state-of-the-art review papers recommending SW technologies. In addition, the review papers for three of the domains (agriculture, rare diseases and Alzheimer's disease) proposed the adoption of a standard data format for data interchange. However, DS, FHIR and GDS solution approaches can be limited to datasets within the same subdomain. Further interlinking these datasets with other existing datasets might not be possible without additional mapping. The mapping would have to transform the data to an adequate data format and establish correspondences between similar entities. SW approaches already provide a standard stack of Web technologies to address this limitation to create Linked Data, a collection of interrelated datasets on the Web. Therefore, SW can effectively address the structure (SR) and semantic heterogeneity (SH) aspects to achieve higher levels of data interoperability.

Machine Learning (ML) and Deep Learning (DL) approaches are also recommended solutions for *Challenge 1: Data Interoperability* in health subdomains like biological data, precision medicine, healthcare and biomedical data science. These approaches have the benefit of being able to extract knowledge from unstructured data with increasing precision and recall, performing the task similar to a human. However, ML (and DL) require a large data volume to build a model for specific data integration tasks. The choice of using ML or SW technologies will depend on the use case but these approaches can be combined to benefit from the advances in both fields [Hogan, 2020]. For example, if datasets are made highly interoperable with semantic context, analysing this data with ML methods will be enhanced in terms of less amount of time needed to train the model and the reduced likelihood of producing errors as the data would be of higher quality.

Finally, the usability of SW and KG for domain experts remains a limitation for the adoption of these technologies when making heterogeneous datasets interoperable. Almost all of the state-of-the-art *review papers* that suggested a SW-KG approach included Domain Expert Usability (DEU) as an existing challenge aspect (14 out of 17, Table 3.2 and 3.4). This finding from analysis of the state-of-the-art review papers support the conclusion of the review from Hogan et al. 2020, which highlights the issue of *"usability of Semantic Web technologies and their accessibility to newcomers"* [Hogan, 2020], after reviewing two decades of research literature. Of the remainder of the main challenge aspects identified in the previous sections (Section 3.1.1 and 3.1.2), Data Quality (DQ) and Accessibility and Availability (AA), are also related to DEU. In order to address these aspects, SW and KG approaches need to be tailored towards addressing researchers' requirements in a particular context so the data is fit for purpose (DQ) (9 out of 17, Table 3.2 and 3.4). The SW and KG approaches can also benefit from a User Centred Design (UCD) [Jokela et al., 2003] approach to refine solutions to focus on achieving the user requirements [Schaaf et al., 2020]. Furthermore, if data is made interoperable at the level that it can be included in researcher's workflows and provided with enough provenance metadata that makes it usable, it can also address the AA challenge aspect, which was present in almost half of the *review papers* related to SW (8 out of 17, Table 3.2 and 3.4).

3.2 Challenge 2: KG usability

This section overviews the state-of-the-art in design tendencies and empirical usability evaluations of User Interfaces (UI) to facilitate the use of standards-based KG approaches (i.e. RDF, OWL or SPARQL) for researchers without practical experience in using the technologies. It focuses on analysing UI approaches where users interact with KG based applications in general (Section 3.2.1) and rare disease and environmental research domains (Section 3.2.2). It then presents an analysis of usability studies in the health domain where users interact with KG based applications (Section 3.2.3). Section 3.2.4 then discusses the key findings arising from the analyses presented.

Making standards-based KG approaches usable depends on the type of users, tasks or goals to achieve and context where these tasks take place. The ISO defines usability as follows:

Usability: *"Extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* – ISO 9241-11:2018

Usability studies need to clearly define these usability aspects as the results, positive or negative, will be indicative that the tool is usable only for the particular group of users in achieving a defined task in a given context. That is one of the reasons why taking a User Centred Design (UCD) approach is considered best practices in usability [Lewis, 2014]. UCD is an iterative process where designers and users collaborate in an active manner towards understanding the context of use and task requirements to assign the adequate interactions between the users and technology [Jokela et al., 2003]. Therefore, usability testing presents an opportunity to promote collaboration between domain experts and computer scientists when developing tools and processes using standards-based KG solutions [Oldman and Tanase, 2018]. The implementation of a UCD process in the development and evaluation of SERDIF is presented in Chapter 5.

The type of users described in this section and the overall thesis follow the Linked Data categorisation for users from Dadzie et al. 2011 [Dadzie and Rowe, 2011].

Lay users. *"(mainstream) Users who do not necessarily understand the intricacies of RDF and other Semantic Web technologies. Such users are computer literate and are able to find information through online resources such as Wikipedia or search engines. Lay users will span the categories of novice to casual users, and while they may have an interest in the data they explore, only a fraction will have in-depth domain knowledge."*

Tech users. *"Expert users who understand SW and other advanced technologies, have experience in using RDF as a data format, and are able to interpret an ontological model."*

Expert users (or domain experts). *"Such users may not necessarily have (expert) knowledge of SW technologies, but are likely to make use of sophisticated, domain specific analysis tools to manage and interact with often very large amounts of complex, heterogeneous data. They are therefore likely to have a very good understanding of data structure and content in their domain, and bring this knowledge to guide both exploratory knowledge discovery and directed information retrieval, to enhance their ability to obtain the insight brought to bear in decision-making."*

Following the iterative design process from UCD, practitioners can track the effectiveness, efficiency and satisfaction by gathering quantitative metrics (i.e. summative usability). Practitioners are recommended to track the task completion and assists (effectiveness), monitor the time spent per task (efficiency) and use one of the available standard usability questionnaires to measure the satisfaction for the users. The standard aspect of the questionnaire facilitates the analysis of the progress throughout the iterations and enables the comparison with similar tools in the particular domain [Lewis, 2014]. Furthermore, the combination of quantitative metrics with observational findings from the sessions, including the introduction, task completion and debriefing times, can provide a more holistic view of the usability problem to address (i.e. formative usability).

Nevertheless, the qualitative analysis involved in usability studies presents a source of subjectivity that can influence the credibility of the results. Practitioners are recommended to report the usability results and methods in a precise, consistent and exhaustive manner towards minimising the subjectivity of the analysis [Nowell et al., 2017]. This includes being transparent about the provenance of the original data and methods to generate the results presented in the report, while making the data and methods available for other researchers trying to reproduce the results.

In this section, first an overview of the general state of the art in the tendencies when designing a UI to facilitate the use of standards-based KG approaches is presented (Section 3.2.1). This subsection is followed by an overview of a sample of state-of-the-art standards-based KG applications for rare disease and environmental research (Section 3.2.2). Then, scientific literature on empirical usability evaluations for these technologies is reviewed for the health domain in particular (Section 3.2.3), organised based on the definition of usability including the goal, type of users, and context of use. Finally, Section 3.2.4 discusses the findings from the previous two subsections that informed the design and implementation of SERDIF and the usability evaluation approach adopted.

3.2.1 User Interaction with standards-based KGs

This section included scientific literature from searching *review articles* Google Scholar using *usability, RDF and SPARQL* keywords, and by targeting publications in the premier workshop on user interaction topics in the Semantic Web community, Visualization and Interaction for Ontologies and Linked Data (VOILA), co-located with the International Semantic Web conference (ISWC). The review articles and the VOILA publications covered scientific literature in the last seven years from the publication of this thesis (2015-2022), as ISWC co-located the first edition of VOILA in 2015.

Review articles. Common complexities in using standards-based KG approaches for researchers that are not familiar with the underlying technologies include the interaction with RDF graphs. This interaction with RDF graphs includes: providing suitable entry points, exploring of the content within a graph, facilitating the analysis of the data structure, granting the possibility to run queries to retrieve and integrate information from the graphs and allow the users to edit or annotate [Aguiar et al., 2021; Dadzie and Pietriga, 2016]. SPARQL [W3C SPARQL, 2013] is the W3C recommendation for querying RDF graphs enabling users to interact with the data in a precise and expressive manner. However, SPARQL is a complex language in terms of syntax and requires an understanding of the underlying structure of RDF graphs or KGs. This complexity presents an obstacle for the adoption by lay users and domain experts without practical experience in KG [Dadzie and Pietriga, 2016; Hogan, 2020; Kuric et al., 2019; Vega-Gorgojo et al., 2016]. Furthermore, the majority of visualisation solutions are designed to be used by technical users with limited usability for the expert users without expertise in KG technologies [De Santo and Holzer, 2020; Klímek et al., 2019]. Expert users have a deep understanding of the data structure and content in their domain and their current involvement and accessibility to KG and Linked Data is limited to few implementations [De Santo and Holzer, 2020; Klímek et al., 2019].

Visualisation tools that facilitate writing SPARQL queries, named query builders, are being developed by the KG community [Kuric et al., 2019]. Query builders or visual query editors hide the complex syntax of the SPARQL language and limit the queries to valid queries. Query builders can be categorised based on the querying approach into: **form-based**, **graph-based**, **natural language-based** and **facet-based**. Form-based interfaces include user inputs (e.g. textual or dropdowns) where the user can select the components that make a query by steps. Graph-based queries utilise a visual approach to select the parts of the graph that you are interested in to build the query around. Natural language-based query builders allow the users to type a question or request using plain language, which is then interpreted and transformed into a SPARQL query. Facet-based builders facilitate the selection of categories or attributes from a list similar to an archive page.

Table 3.5: Summary of the *review papers* that describe the general state of the art in the tendencies when designing a UI to facilitate the use of standards-based KG approaches

Publication	**Type of review and topic**	**Reviewed work (#)**	**Years**
[Bikakis and Sellis, 2016]	A survey study of Linked Data exploration and visualisation systems	64	2003-2015
[Vega-Gorgojo et al., 2016]	A comparative study evaluating the usability of a graph-based tool against a form-based tool	2	2015-2016
[Dadzie and Pietriga, 2016]	A literature review on Linked Data visualisations including usability and utility evaluation approaches	17	2011-2015
[Kuric et al., 2019]	A comparative usability evaluation of three query builders to explore KGs for lay users	3	2015-2107
[Klímek et al., 2019]	A survey study of tools for Linked Data consumption	16	2013-2017
[Desimoni et al., 2020]	A comparative study of Linked Data visualisation tools with SPARQL endpoint support	10	2010-2020
[De Santo and Holzer, 2020]	A survey study on Linked Data interaction tools including research from the ACM Special Interest Group on Computer-Human Interaction venues	43	2004-2018
[Aguiar et al., 2021]	A survey study on user interaction with Linked Data for users with non-expert users	18	2004-2020

While graph-based query builders represent the majority of visualisations tools and they can work for small graphs, the user may feel overwhelmed as the visualisation becomes intelligible when interacting with large graphs such as Linked Open Data (LOD) [Aguiar et al., 2021; Dadzie and Pietriga, 2016]. KG practitioners should consider more complex queries or advanced approaches to visualise a subset of the large graph that is relevant to the user [Bikakis and Sellis, 2016]. Examples include transforming the raw data in the graph by aggregating the results, clustering groups of results to identify patterns or partitioning a particular view of the large graph.

Natural language-based builders can struggle with complex queries as these need to be interpreted, which leads to tailored applications that are able to assimilate particular queries. Following, facet-based builders expressivity is also limited to filtering concepts in a query (i.e. classes, properties and instances), which may not be enough to conduct certain tasks for users. From the four query builder designs, the form-based query offered the best usability results for lay users trying to explore the LOD [Kuric et al., 2019; Vega-Gorgojo et al., 2022,1]. Form-based queries are easier to learn and navigate, hiding the underlying graph data model.

While visualisation tools can help in interacting with RDF data, the usability of these tools need to be validated by usability evaluations for the target group of users and tasks in a given context. That is because the usability of the visualisation tools vary based on the domain, the type of data and the set of relevant tasks for a user with a particular expertise and technical skillset [Desimoni et al., 2020].

The *review papers* terminology defined in Section 3.1.1 is also used in Table 3.5. for the categorisation of the reviewed research.

VOILA publications. The target review on VOILA publications focused on publications in sessions about visual SPARQL querying and linked data exploration. The initial search identified 41 candidates for the review from all the editions of the VOILA workshop. In case that no explicit name was given to the sessions (2020, 2021 and 2022 editions), all the publications were selected for the initial screening. From the initial set of publications, 11 of them were discarded due to being publications that described ontologies, vocabularies extensions, being *review papers* included in the previous section (Table 3.5) or not including SPARQL queries. The remaining 30 publications were analysed based on the use of form-based query builders and usability evaluation to validate the approach for a type of users (i.e. lay, tech and expert users), and summarised in Table 3.6. The dimensions of the analysis are selected based on the results from the **review papers** section. The summary of the publications follows the following pattern to facilitate the analysis of the results: “A X-based UI for Y users to do Z”, where X is the type of UI, Y one of the three types of users defined in the previous section and Z the goal of the UI. The authors of this thesis used screenshots provided to define the UI type when this element was not explicitly defined. Some of the publications did not specify the type of users that would be using the visualisation and therefore, it was left blank instead of assuming one.

Table 3.6: Summary of the VOILA publications on visual SPARQL querying and linked data exploration. The black circle indicates the presence of the element and the blank circle the absence of it. An asterisk next to the circle indicates that the usability evaluation did not follow best practices in usability.

Publication	**Summary**	**Form-based query**	**Usability evaluation**
[Bottoni and Ceriani, 2015]	A block-based programming tool for lay users to experiment with queries	○	○
[Soylu et al., 2015]	A widget-based UI mashup for lay users to formulate queries based on an ontology	○	○
[Zainab et al., 2015]	A graph-based visual interface for expert users to run federated queries in the general and biomedical domains	○	●*
[Trinh et al., 2015]	An autocomplete input box for lay users to semantically annotate text	○	●*
[Florenzano et al., 2016]	A graph-based visual aid for tech users to query unfamiliar RDF datasets	○	○
[Weise et al., 2016]	A graph-based visual tool for lay users to extract and visualise the schema information of Linked data sources based on VOWL notation	○	○
[Čerāns and Ovčiņņikova, 2016]	A UML-based tool for lay users to define analytic queries	○	○
[Blinkiewicz and Bak, 2016]	A form-based interface for lay users to create R2RML mappings and test the mappings with a graph-based UI with queries against ontologies	●	○
[Khalili and Meroño-Peñuela, 2017]	A facet-based UI for lay users to explore interlinked datasets	○	○
[Fuenmayor et al., 2017]	A facet-based UI for lay users to explore multiple RDF KGs	○	○

[Mitschick et al., 2017]	A facet-based query interface for lay users to enable text and semantic searching	○	●*
[Čerāns et al., 2017]	A UML-based tool for expert users to construct complex queries	○	●*
[Regalia et al., 2017]	A form-based Web interface for tech users to explore remote RDF datasets via public endpoints	●	○
[Leskinen et al., 2018]	A faced-based UI for lay users to create data-analytic visualisations from the results of a SPARQL endpoint	○	○
[Bartolomeo et al., 2018]	A graph-based UI for lay users to build queries based on GRAPHOL ontologies	○	○
[Křemen et al., 2018]	A form-based UI for tech users to explore SPARQL endpoints	●	●*
[Tartari and Hogan, 2018]	A form-based UI for tech users to find the shortest path between two nodes	●	●*
[Graux et al., 2020]	A facet-based UI for tech users to visualise in real-time changes in Wikidata	○	○
[Kulahcioglu et al., 2020]	A graph-based UI for expert users to visually analyse log graph data in industrial equipments	○	●*
[Ramadhana et al., 2020]	A facet-based UI for expert users to analyse knowledge imbalance in Wikidata	○	●
[Navarro-Gallinad et al., 2020]	A form-based UI for expert users to link health and environmental data	●	●
[Parra and Hogan, 2021]	A text-based UI for tech users to autocomplete queries based on a graph summary	○	○
[Raissya et al., 2021]	A wide range of visualisations for tech users to assist them in extracting patterns and insights from query results over KGs	○	○

[Graux et al., 2021]	A facet-based UI for lay users to explore temporal information available in SPARQL endpoints such as Wikidata	○	○
[Ehrhart et al., 2021]	A facet-based UI for expert users to explore information in a Knowledge Graph	○	●
[Yacoubi et al., 2022]	A visual exploration of RDF data cubes for lay users to view raw and aggregated weather linked data	○	○
[Čerāns et al., 2022]	A UML-based tool for tech users to query DBpedia	○	○
[Bruyat et al., 2022]	An autocomplete input box based for tech users to edit RDF documents based on RDFS and/or SHACL schemas	○	●*
[Haller et al., 2022]	A facet-based UI for lay users to enable exploratory searches on manufacturing KGs	○	●*
[Serrano et al., 2022]	A graph-based tool for expert users to visualise semantic post-hoc explanations for predictions made by AI models	○	●*

The reviewed VOILA publications opted for facet- (9 out of 30 Table 3.6) and graph-based (9 out of 30 Table 3.6, including UML-based interfaces) UIs to facilitate and/or enable the user interaction with standard-based KGs, supporting the finding from the **review papers** section. Form-based UIs were present to a lesser extent throughout 2015-2022 (5 out of 30 Table 3.6) without a trend indicating the recent preference for this type of UIs.

Regarding usability evaluations, almost half of the VOILA publications included them as part of their tools evaluation (13 out of 30 Table 3.6). The number of publications with usability evaluations remained more or less constant at almost 2 per year with an increase of one unit in the 2020 and 2022 editions. However, only a few publications conducted usability studies following best practices (3 out of 13 Table 3.6) as the rest limited the evaluation to quantitative or qualitative metrics only, or technical-users instead of the intended user type; or the description provided for the evaluation was not enough to identify if best practices were being used. The importance of the tools being usable beyond tech users was also mentioned in the future work sections of the reviewed papers. This is reflected in the table with the most common type of end-user being lay users (14 out of 30 Table 3.6), and then, tech (9 out of 30 Table 3.6) and expert users (7 out of 30 Table 3.6) with less frequency. The type of users distribution remained similar throughout all editions with the first three (2015, 2016 and 2017) having a tendency towards lay users.

3.2.2 User Interaction with standards-based KGs in the rare disease and environmental domains

This section overviews a sample of state-of-the-art standards-based KG applications for rare disease and environmental research. This section includes scientific publications that were not published in VOILA but searched in Google Scholar using the *usability, RDF, SPARQL, health, rare disease and environmental data* keywords. This sampling review focuses on the type of data sources integrated in the KG, the requirement of the users to be tech users to use the applications and whether a usability test to evaluate the effectiveness and usefulness of the approach has been undertaken.

Rare disease research. Roos et al. 2017 discuss the impact of applying standards-based KG 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. Researchers could collaborate by designing a visual interface from a domain expert's requirements that benefits from the underlying technologies for an easy and meaningful access to the combined data.

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; Roos et al., 2016; 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 [Choquet et al., 2015] 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 an evaluation to assess the usability and potential usefulness of the visual interface for domain experts.

In the same biomedical domain, other approaches require users to have 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,1], 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 customised approach to evaluate the visual interface impeding the comparison with similar visualisation tools.

Environmental research. Earth observation studies, a type of environmental data gathered with remote sensing technologies, have been exploring the application of standards-based KGs to facilitate the interoperability of the data sources towards an effective use of data for analysis. Towards this end, the H2020 project DeepCube adopted RDF to describe Copernicus data such as optical, climate, meteorological, industry and social media data as RDF, with the goal of making it available as Linked Data on the Web [Gervasi et al., 2021]. The technologies used in the project included the Earth Observation Data cubes [Mahecha et al., 2020], a data cube with latitude, longitude, time and variable dimensions developed to facilitate the access and use of multivariate datasets; and the Semantic Earth Observation Data cubes [Augustin et al., 2019; Giuliani et al., 2019] to enrich the cubes with enough metadata for an adequate interpretation of the cube data by humans and machines. Furthermore, DeepCube extended the Sextant platform [Nikolaou et al., 2015], which is a web-based platform to interact with linked geospatial data in a user-friendly and visual manner. However, no usability study was conducted to support that the Sextant (and the extension) was indeed usable for expert and lay users.

Other interfaces that benefit from uplifting geospatial and temporal datasets to RDF have been developed for weather, climate, environmental intelligence and forest exploration applications. Atemezing et al. 2013 [Atemezing et al., 2013] describe the uplift process of public meteorological data from the Spanish meteorological office available as Linked Data. The authors designed a facet-based interface to explore the meteorological data, where technical-, expert and lay users could interact with the data by clicking geographical locations on a map. In a similar manner, Roussey et al. 2020 uplift the data from a weather station in France to RDF [Roussey et al., 2020]. The authors designed a SPARQL interface with some sample queries to guide technical-users in querying the weather data and visualise the results as a time series plot. Regarding climate applications, the Australian Bureau of Meteorology published an RDF dataset, ACORN-SAT, which included homogenised daily temperature observations for 112 locations throughout Australia for the last 100 years [Lefort et al., 2017]. The interaction with the linked temperature data was facilitated by a simple form-based interface, where the user can select a day range for a given location with a dropdown box.

Regarding environmental intelligence, Janowicz et al. 2022 developed a cross-domain Knowledge Graph (KG) and geoenrichment services for expert users to access spatial and temporal data for any location on Earth, called KnowWhereGraph [Janowicz et al., 2022]. This KG pre-integrated data from events, administrative boundaries, soils, crops, climate, transport and for answering specific questions. On top of the KnowWhereGraph, a UI was developed towards facilitating the use for a specific type of expert user including geographic information systems specialists, disaster relief specialists and decision-makers assessing the impact of ongoing wildfires. While the KG was meant to be usable for particular expert users, no usability evaluation was conducted not included in the future plans for the KnowWhereGraph. That was also the case for the weather and climate studies reviewed where usability of the standard-based KG approach was not considered, limiting the evidence for the future adoption of the approaches. The usability aspect of these technologies was only considered by Vega-Gorgojo et al. 2022 for the forest explorer tool [Vega-Gorgojo et al., 2022]. This tool facilitated the exploration of the cross-forest dataset, which describes the forestry inventory and land cover map from Spain, for expert and lay users. The usability comprised a profile of the users, a dataset assessment, the completion of a standard post-survey questionnaire and a recording of user feedback.

Based on the rare disease and environmental studies overviewed in this section, user interfaces are being used to promote the interaction with KG for expert users in these domains. However, the limited presence of usability studies (2 out of 15 reviewed studies) may diminish the adoption of these technologies by expert and lay users, and even for tech users; as no evidence is provided that they are usable by the intended user. In addition, the data sources integrated in the rare disease studies had a predominant focus on biomedical data and did not consider the linkage between health and environmental data. For environmental studies, the integration of data was extended to different types of data from weather, climate and other sorts of spatial data through spatiotemporal queries.

3.2.3 Usability evaluations for standards-based KGs in the Health Domain

This section overviews the state of the art in empirical usability evaluations of UIs that hide the complexities of using standard KG approaches for researchers in the health domain. This section included scientific literature from searching journal and conference publications on Google Scholar using *usability evaluation, RDF, SPARQL and health* keywords. The publications covered scientific literature in the last seven years from the publication of this thesis (2015-2022) in line with the other sections in this chapter.

The review is centred on eight aspects important to analyse usability studies based on the definition of usability (ISO 9241-11:2018), common best practices [Jokela et al., 2003; Lewis, 2014] and the importance of reproducibility to increase the traceability and verification of the data analysis [Nowell et al., 2017]. These aspects are the (1) goal of the UI in (2) a particular use case, (3) the type of end-users or participants included in the usability evaluation (i.e. lay, tech and expert users), (4) the user-centred design for the UI, (5) number and type of phases, (6) the inclusion and analysis of quantitative and qualitative data, (7) the use of a standard post-survey questionnaire, and (8) the availability of the raw and processed data and the sufficient description to reproduce the results of the study (Tables 3.7 and 3.8). In addition, the authors of this thesis considered the need for a training or tutorial for the participants before the usability experiment as the ninth aspect (9) to account for in usability studies since it was a common factor across the publications reviewed in this section. The work of this thesis is provided as the last row in Table 3.7 and 3.8 for comparison with the reviewed studies.

The analysis of the state of the art includes a variety of use cases related to the health domain, representing the overall progress and willingness of making standards-based KGs usable within the health domain in recent years (2015-2022). Most of the tools to facilitate the use of standards-based KGs (7 out of 9, Table 3.7) focused on the data exploration and integration tasks in the health domain, reinforcing the importance and complexity of addressing the *Challenge 1: Data interoperability* and in making these tasks more efficient. The remaining tools had the objective of promoting the data annotation and retrieval tasks in researchers' workflows.

A common approach when evaluating the usability of a UI is the combination of quantitative and qualitative metrics (QQ), identifying the issues and the underlying reason why they are an issue for the users respectively (4 out of 9, Table 3.8). The qualitative metrics concur with the ones described in Chapter 5 of this thesis, being the think aloud thoughts recorded on the transcriptions of the usability sessions, notes taken and open comments from the usability surveys. Thematic analysis was the preferred method to analyse the qualitative data. In addition, two studies from the same authors conducted a heuristic evaluation to map identified usability issues to usability heuristics [He et al., 2020,1]. Nevertheless, the number of evaluators for the data analysis step was not always explicit in the analysis nor in the author's contribution section of the work. Mentioning the evaluators number and their role in the analysis is recommended to minimise the subjectivity of the qualitative analysis [Nowell et al., 2017].

Table 3.7: Summary of key features usability aspects related to the definition of usability in empirical usability evaluations of standard KG tools in the health domain. The black circle indicates the presence of the element and the blank circle the absence of it.

Publication	Summary	Goal	Use case	Expert users
[Minutolo et al., 2022]	A conversational agent for querying Italian Patient Information Leaflets and improving health literacy	Data exploration	Patient information	○
[He et al., 2019]	An interactive graph-based visualisation for dietary supplement knowledge graph	Data integration and exploration	Dietary supplement	○
[He et al., 2020]	A web-based crowdsourcing-integrated semantic text annotation tool for building a mental health knowledge base	Data annotation and extraction	Mental health	○
[Hu et al., 2017]	A semantic search engine for Bio2RDF	Data exploration	Life sciences	●
[Ali et al., 2022]	A Linked Open Data system for structuring and transforming COVID-19 data	Data integration and exploration	COVID-19	●
[Dafli et al., 2015]	An extension of the OpenLabyrinth virtual patient authoring and deployment platform for the repurposing and retrieval of existing virtual patient material	Data repurposing and retrieval	Health education	●
[Stöhr et al., 2021]	A web-based metadata management app that is usable for scientists to compile and use rich metadata	Data integration	Lung research	●
[Marcilly et al., 2020]	A comparative study that compares three approaches for supporting the use of MedDRA by pharmacovigilance specialists	Data exploration	Pharmacovigilance	●
[Hanlon et al., 2021]	A storyboard for health professionals to explore, integrate and assess the quality of health data	Data exploration, quality and integration	Medical	●
[Navarro-Gallinad et al., 2022]	A framework for health data researchers to link health events and environmental data	Data integration	Rare diseases	●

Table 3.8: Summary of key features related to common best practices and traceability of the data analysis in empirical usability evaluations of standard KG tools in the health domain. The efficacy and quality of the system algorithm are not included in this summary. Acronyms: Follow-up evaluation (>>), Different metrics in each phase (ab), Standard Post usability Survey (SPS), User Centred Design (UCD), No Training or Tutorial before tasks (NTT), Qualitative and Quantitative usability metrics analysis (QQ), Reproducible Results (RR). The black circle indicates the presence of the element and the blank circle the absence of it.

Publication	**UCD**	**Phases (type)**	**QQ**	**SPS**	**NTT**	**RR**
[Minutolo et al., 2022]	○	2 (ab)	○	●	○	○
[He et al., 2019]	●	3 (>>)	●	●	○	○
[He et al., 2020]	●	4 (>>)	●	●	○	○
[Hu et al., 2017]	○	1	○	○	○	○
[Ali et al., 2022]	○	1	○	●	○	○
[Dafli et al., 2015]	○	2 (>>)	○	●	○	○
[Stöhr et al., 2021]	○	1	●	●	○	○
[Marcilly et al., 2020]	●	2 (ab)	○	●	○	○
[Hanlon et al., 2021]	●	1	●	●	●	○
[Navarro-Gallinad et al., 2022]	●	3 (>>)	●	●	●	●

Quantitative metrics included Standard Post usability Surveys (SPS) allowing researchers to compare the results with further versions or similar tools. Most of the studies reviewed (8 out of 9) used the System Usability Scale (SUS) questionnaire [Brooke, 1996] to gather quantitative data about the usability and efficacy of a tool for a specific purpose [Ali et al., 2022; Dafli et al., 2015; He et al., 2020,1; Hu et al., 2017; Marcilly et al., 2020; Minutolo et al., 2022; Stöhr et al., 2021]. The remaining study [Hanlon et al., 2021] used the Post-Study System Usability Questionnaire (PSSUQ) as the usability questionnaire [Lewis, 2002]. The choice between the use of SUS or PSSUQ depends on each case. For example, the evaluation approach presented in this thesis decided to adopt the PSSUQ since the questionnaire was designed specifically for scenario-based usability studies. The PSSUQ also provides scores for the System Usefulness, Information Quality, and Interface Quality scales for a more detailed view of how the HDRs perceived the usability of the SERDIF framework. In addition, the standard questionnaires have been validated to be sensible to small sample sizes (<10), which tends to be a common size for the reviewed studies, as well as, including an objective and quantitative view to the study. More detail is provided in Chapter 5.

Regarding the evaluation strategy, most of the reviewed studies included health experts from professional and researcher profiles. The purpose of the studies was to facilitate the use of SW technologies by domain experts. Student participants were also included for convenience in four of the studies [Dafli et al., 2015; He et al., 2020,1; Hu et al., 2017]. While most of the studies planned for multiple phased approaches to refine the tools, some mentioned the intention to do further evaluations in the strategy or plan as future work since the studies were not at that stage (Table 3.8). Thus, a multi-phased approach can be considered as best practice in the health domain. A UCD approach was followed by only four of the reviewed studies (Table 3.8), even though it is considered best practice for usability studies.

Only three studies provided follow-up evaluations comparing the SUS results of experts, college students, consumers or workers when using a UI (Table 3.8). While the usability improved in [Dafli et al., 2015], the authors from [He et al., 2020,1] argue that the SUS results did not follow an increasing trend from phase to phase (i.e. improving the usability) since new functionalities or features requested by the users were added, making the UI more complex. Coinciding with the evaluation results of SERDIF in Chapter 5, they concluded that dividing complex tasks in subtasks and choosing simplicity over including all functionalities will improve the usability of the UI. Nonetheless, only [Hanlon et al., 2021] provided information about the assistance required by the usability moderator for the participants when completing the tasks. The assistance combined with the completion or failure assessment and the time per task can be a key combination of metrics when trying to understand the adequacy of a task in the experiment and the complexity for the user. Furthermore, the reviewed studies included a tutorial or training step at the start of the experiment, reducing the first barrier of complexity when using the tool (Table 3.8).

The reviewed studies did not provide either the raw and processed data or sufficient description to reproduce the results of the study as supplementary information or deposited in an open data repository generating a persistent Digital Object Identifier (DOI). The additional information provided was limited to tables and figures linking the quantitative metrics, themes, usability findings and recommendations. This could complicate the reproducibility of the work (e.g. the qualitative part) by other researchers and make it hard for researchers, who are novice in usability studies but knowledgeable in SW and KG technologies, to learn how to analyse the results.

3.2.4 Discussion

Visualisation tools can facilitate the use of standard KG approaches in terms of integrating, exploring, retrieving, editing or annotating the data in the RDF graphs. Form-based query builders are the recommended UI design by *review papers* included in the review for the general domain (Section 3.2.1). The selected scientific literature reviewed represents only the academic literature but an important amount of eHealth apps, not necessarily using standard-based KG approaches, is being developed by developers in the industry [Maramba et al., 2019]. The industry sector does not usually publish usability studies to protect their technological or usability expertise advantage over their competitors.

However, usability still remains a problem for the adoption of KGs. Further usability studies are needed for diverse use cases as usability relies on conducting a series of tasks by a particular group of users under a defined context of use. Therefore, answering the call for more research related to reducing the expertise required to benefit from the technologies and engaging new researchers from other domains [Hogan, 2020].

None of the reviewed empirical usability evaluations studied usability of KGs in the context of rare disease research nor health-environment linking applications of focus for the research of this thesis (Section 3.2.2). While these contexts might have similarities with some of the reviewed health studies, such as types of data and domain experts involved, differences are also present. Examples of these differences include the increased importance of integrating multiple data sources, and the adequacy of the linked data to conduct specific data analysis in line with the sparsity and limited amount of data available given the rarity of the health condition [Haendel et al., 2020; Piel et al., 2020]. While there are standard KG approaches being used to address the interoperability challenges in rare disease and environmental research, the usability of the user interfaces to promote the adoption for lay, tech and expert users was not studied (Section 3.2.2). Furthermore, application of standards-based KG approaches for rare disease research was limited mainly to biomedical and clinical data without including environmental risk factors.

While UCD presents an opportunity to establish the collaboration channels between researchers and computer scientists, only half of the reviewed empirical usability evaluations (4 out of 9, Table 3.8) in the health adopted a UCD. Furthermore, a similar tendency is observed regarding the rest of usability best practices (Table 3.8). The reviewed studies in Section 3.2.3 lacked sufficient documentation in terms of the raw and processed data, and methods to reproduce the results of the study. This documentation could have improved the credibility of the results encouraging future researchers to follow the designs and approaches that were found usable, instead of re-designing a solution from scratch. In addition, the results and the description of the methods can guide and promote future usability studies of tools to make the KG more accessible to domain experts, and researchers seeking to learn about SW technologies. That is why, this thesis details step by step the adopted usability approach in Chapter 5.

3.3 Preliminary study

This section outlines a preliminary usability study on a prototype user interface for Health Data Researchers (HDR) to retrieve environmental data linked to clinical data. This usability study was informed and served as a validation of the design choices from the state-of-the-art findings as presented in Section 3.1 and 3.2. The content of this section has been peer-reviewed, presented and published [Navarro-Gallinad et al., 2020] in the 5th International Workshop on Visualization and Interaction for Ontologies and Linked Data (VOILA) 2020, co-located with the International Semantic Web conference (ISWC) 2020. VOILA is the premier workshop on user interaction topics in the Semantic Web community. The content has been adapted and summarised to follow the structure for a UCD [Jokela et al., 2003]: understand context of use (Section 3.3.1), identify initial expert user requirements (Section 3.3.2), design a prototype user interface (Section 3.3.3) and evaluate against the requirements (Section 3.3.4).

3.3.1 Understand context of use

HDRs face technical challenges when integrating data from different sources to generate new insights into the underlying disease mechanism (Section 3.1.1). When addressing the data integration challenges with emerging technologies like KGs, researchers typically require knowledge engineers to access, explore and retrieve data that they are interested in from datasets (Section 3.2.1).

Two example healthcare data linkage projects that use KGs to integrate their diverse datasets are the HEalth data LInkage for ClinicAL benefit (HELICAL) project [HELICAL, 2023] and the AVERT project (AVERT) [AVERT, 2022; Reddy et al., 2019]. Researchers in both projects are studying the environmental triggers of Anti-neutrophil cytoplasm antibody (ANCA)-associated vasculitis (AAV) flares[1] and its prevention. AAV is a rare autoimmune disease of unknown aetiology with flares that can progress into damage to vital organs [Kitching et al., 2020]. The current consensus is that this aetiology involves a complex interaction between environmental and epigenetic factors, in a genetically susceptible individual [Kitching et al., 2020]. Identifying environmental triggers of AAV can lead to flare prediction models for a precision medicine approach for people with this condition.

The context of use is further described in Section 5.2.1.

3.3.2 Identify initial expert user requirements

An initial set of requirements were gathered from expert researcher meetings and through undertaking a consensus process with HDRs from the AVERT and HELICAL

[1] sudden appearance or worsening of the symptoms of a disease or condition

projects.

The author of this thesis joined both research projects that acted as focus groups, and actively participated in weekly meetings to understand the nature of the tasks and the context of use (Section 3.3.1). During the first three months the author studied documentation related to the context of use, including the health and KG background necessary to design a solution. After studying the documentation and recording the findings from assisting the research meetings, the author and supervisors of this thesis distilled the initial expert user requirements. The list of the expert requirements was presented as part of an oral presentation in one of the research meetings, where HDRs from AVERT and HELICAL were present. The list was then modified to achieve consensus from the experts within the context of use described in Section Section 3.3.1.

An effective solution would thus intend to achieve the following expert user requirements:

Requirement 1 (R1.0). *Enable HDRs to query specific clinical patient data to retrieve linked environmental data, without the need for knowledge of the underpinning semantic web technologies.*

Requirement 2 (R2.0). *Support the understanding of the use and limitations of the linked environmental data to support identification of flares for rare diseases.*

Requirement 3 (R3.0). *Export selected clinical and environmental data to be used as input in statistical models for data analysis.*

3.3.3 Design a prototype user interface

A prototype User Interface (UI) was designed to address the expert user requirements presented in Section 3.3.2. The UI operates on top of an existing KG from the AVERT project [Reddy et al., 2019] that contains health and environmental data. The health data includes 2.6M triples of infectious disease and clinical data from people with AAV data in Ireland from approximately early 2000's to the present. The environmental data consists of 27.5M triples of weather and air pollution data for the same location and period.

The UI was structured in two panels, the query and main panels (Fig. 3.1).

Query panel. This panel granted a non-tech user the option to select different clinical parameters related to flares and how to link them with environmental data (Fig. 3.1A). The parameters selected by the user were then substituted into a SPARQL query template, URL encoded and executed against the data in the triplestore. This panel is aimed towards satisfying R1.0.

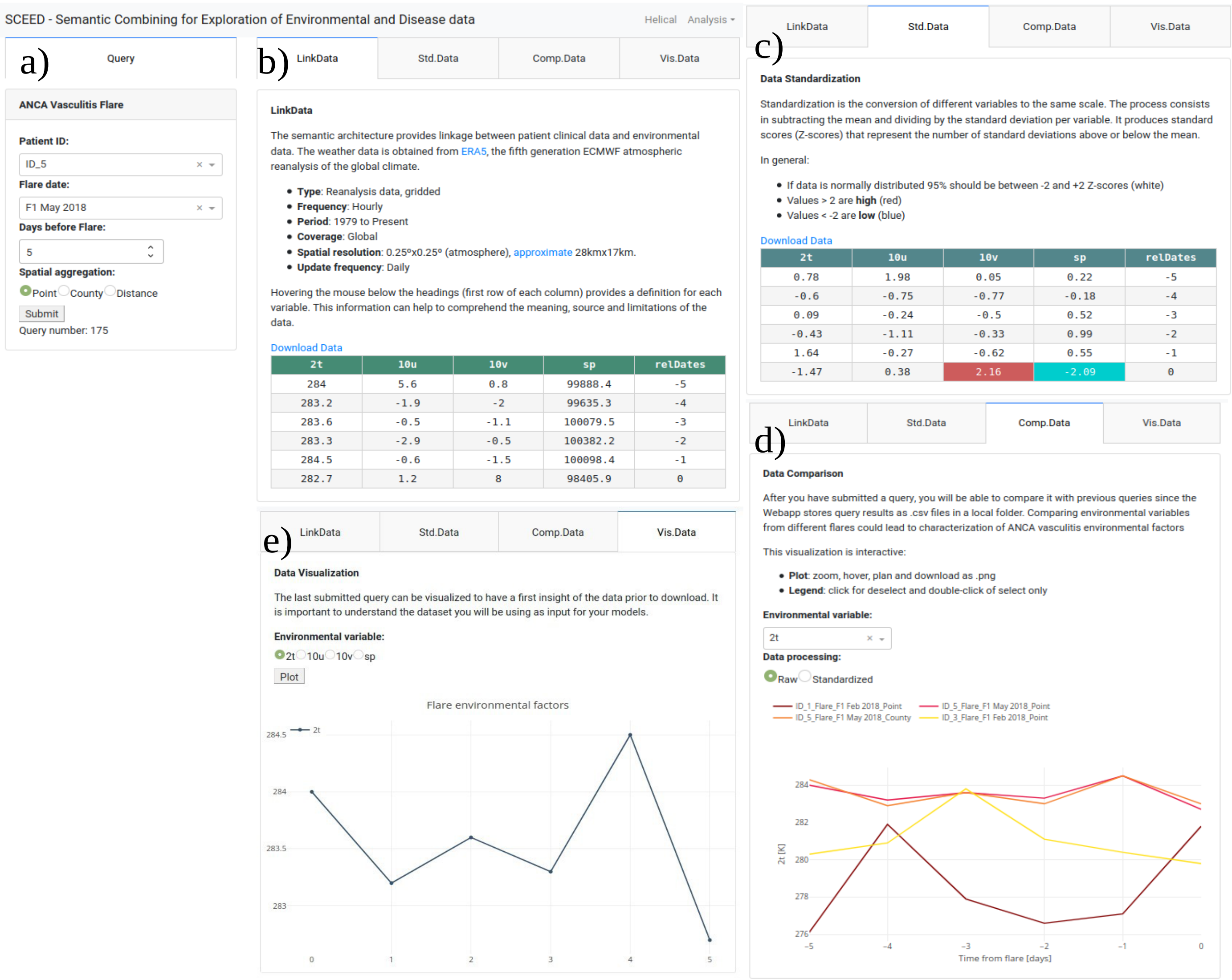
Main panel. This panel is divided into 4 tabs: link data, standard data, comparative data and visualisation. After submitting a query, the user got the environmental data linked to the clinical patient data as an interactive data table in the link data tab. The data table can be exported as a CSV file to support R3.0. The standard data tab granted the user the possibility to compare the environmental variables, which have been (statistically) standardised to the same scale using Z-scores, in a data table (Fig. 3.1B). The standard data table had some highlighted values with colour encoding depending on the category of the value (red for high and blue for low as in Fig. 3.1C) and the standard values were also exportable as a CSV file. The comparative data tab facilitated the comparison of previous queries with the current one. The user was presented with a multi-line interactive plot to explore trends, seasonality, comparison and check outliers (with the standard view) to improve their understanding of the environmental data previous to the patient's flare event in support of R.2.0 (Fig. 3.1C). The visualisation tab provided the user with a cleaner view of the current query results for a quick insight of the data prior to the export. Thus making sure the data was usable and appropriate to support a specific research question, also in support of R.2.0 (Fig. 3.1D).

The UI was coded in *Python* (v3.6) using the *Dash* (v1.7.0) package as a framework that facilitates building cross-platform analytical platforms. The UI was coded dynamically, displaying the data available as it is imported into the triplestore by querying the endpoint. This approach was efficient in this context as both clinical and environmental data was constantly being updated, the data collection process in the research workflow was an ongoing process.

3.3.4 Evaluate against the requirements

A usability experiment was then undertaken on the prototype UI with HDRs from the AVERT and HELICAL projects with the goal of testing the idea of using a UI and gathering real feedback early on the development process to satisfy the three initial requirements presented in Section 3.3.2.

The experiment was structured in three parts: a brief introduction to the UI background, a series of tasks to be completed by the participants and a follow up post-questionnaire. The tasks were designed with real workflows in mind and selected to assess the achievement of the three initial requirements. The participants were asked to complete the tasks following a think-aloud protocol [Boren and Ramey, 2000], which encouraged participants to speak their thoughts as they completed the session. The sample size included 7 expert participants, which were researchers in the AVERT and HELICAL projects without practical experience in KG, but that could benefit from a KG approach for their studies.



SCEED - Semantic Combining for Exploration of Environmental and Disease data

ANCA Vasculitis Flare

Patient ID: ID_5

Flare date: F1 May 2018

Days before Flare: 5

Spatial aggregation: ● Point ○ County ○ Distance

Submit

Query number: 175

LinkData

The semantic architecture provides linkage between patient clinical data and environmental data. The weather data is obtained from ERA5, the fifth generation ECMWF atmospheric reanalysis of the global climate.

- **Type**: Reanalysis data, gridded
- **Frequency**: Hourly
- **Period**: 1979 to Present
- **Coverage**: Global
- **Spatial resolution**: 0.25°x0.25° (atmosphere), approximate 28kmx17km.
- **Update frequency**: Daily

Hovering the mouse below the headings (first row of each column) provides a definition for each variable. This information can help to comprehend the meaning, source and limitations of the data.

Download Data

2t	10u	10v	sp	relDates
284	5.6	0.8	99888.4	-5
283.2	-1.9	-2	99635.3	-4
283.6	-0.5	-1.1	100079.5	-3
283.3	-2.9	-0.5	100382.2	-2
284.5	-0.6	-1.5	100098.4	-1
282.7	1.2	8	98405.9	0

Data Standardization

Standardization is the conversion of different variables to the same scale. The process consists in subtracting the mean and dividing by the standard deviation per variable. It produces standard scores (Z-scores) that represent the number of standard deviations above or below the mean.

In general:

- If data is normally distributed 95% should be between -2 and +2 Z-scores (white)
- Values > 2 are **high** (red)
- Values < -2 are **low** (blue)

Download Data

2t	10u	10v	sp	relDates
0.78	1.98	0.05	0.22	-5
-0.6	-0.75	-0.77	-0.18	-4
0.09	-0.24	-0.5	0.52	-3
-0.43	-1.11	-0.33	0.99	-2
1.64	-0.27	-0.62	0.55	-1
-1.47	0.38	2.16	-2.09	0

Data Comparison

After you have submitted a query, you will be able to compare it with previous queries since the Webapp stores query results as .csv files in a local folder. Comparing environmental variables from different flares could lead to characterization of ANCA vasculitis environmental factors

This visualization is interactive:

- **Plot**: zoom, hover, plan and download as .png
- **Legend**: click for deselect and double-click of select only

Environmental variable: 2t

Data processing: ● Raw ○ Standardized

ID_1_Flare_F1 Feb 2018_Point
ID_5_Flare_F1 May 2018_Point
ID_5_Flare_F1 May 2018_County
ID_3_Flare_F1 Feb 2018_Point
2t [K]
284
282
280
278
276
-5
-4
-3
-2
-1
0
Time from flare [days]

Data Visualization

The last submitted query can be visualized to have a first insight of the data prior to download. It is important to understand the dataset you will be using as input for your models.

Environmental variable: ● 2t ○ 10u ○ 10v ○ sp

Plot

Flare environmental factors

Figure 3.1: Prototype UI with multiple views after finishing the tasks from the study. a) Query section. The user can select an option from a dropdown display list for Patient ID and Flare date, a numeric value for Days before Flare and an input from the radio buttons in Spatial aggregation. Data tabs: b) Link Data, c) Std.Data, d) Comp.Data and e) Vis.Data; the different tabs allow the user to navigate, compare, visualise and export meaningful information. Each tab starts with an introductory text information to guide the user and ends with a data visualisation as table or graph.

During the experiment, quantitative and qualitative usability metrics were gathered to support the findings of the experiment, including completion of the task, time per task and the observational findings from the session transcripts. At the end of the experiment, the participants were asked to complete a PSSUQ [Lewis, 2002], a standard quantitative evaluation instrument for the satisfaction aspect of the usability given a context and set of participants. The usability metrics are explained in detail in Section 5.1.4.

The results obtained from the usability experiment indicated that the UI was an adequate initial design to fulfil the expert requirements (Section 3.3.2). However, new features would be necessary for comprehending the use and limitations of the environmental data for rare disease flare discovery.

The evaluation approach was able to highlight the successful aspects of the UI, identify the items that needed improvement and the new features to be added. However, a more comprehensive and systematic approach would be required to analyse the observational findings from the session transcripts. The findings will be analysed following the six steps of a thematic analysis described by Braun and Clarke 2006 [Braun and Clarke, 2006] to provide further evidence for the changes and claims made (Chapter 5).

3.4 Summary findings

This chapter reviewed the state of the art on existing approaches and recommendations to make the data integration process effective by making data interoperable, and how to make this process usable for scientific research. The findings of this chapter are summarised below.

Challenge 1: data interoperability. A standard KG approach based on SW is the proposed solution for *Challenge 1: data interoperability* but this solution needs to be usable for domain experts in a given context to achieve a set goal.

Challenge 2: KG usability. A form-based query builder is the preferred solution for *Challenge 2: KG usability* but this solution needs to be validated by an empirical usability evaluation following best practices.

The findings from these analyses were initially explored and evaluated in a preliminary study that undertook an initial exploration of the usability of a standards-based KG approach to facilitate the data integration tasks for a group of researchers studying the health outcomes associated with a rare disease. A key finding from this preliminary study was that although there was promising initial results, that a more comprehensive and systematic approach needed to be designed in order for the solution to be able to cope with a wider diversity of data, domain experts, and use cases related to the linkage of health event related data with environmental data. In the next chapter (Chapter 4) the design and design considerations of the resultant framework called SERDIF is discussed. Chapter 5 then presents the iterative implementation and evaluation of the framework.

Chapter 4

Design

This chapter describes the design of the Semantic Environmental and Rare Disease data Integration Framework (SERDIF). Implementation details are presented as part of Chapter 5. The framework has been developed based on Health Data Researcher (HDR) requirements for addressing data interoperability challenges in health research that would benefit from applying a Knowledge Graph (KG) approach. The expert requirements were derived using a User-Centred Design (UCD) approach (Section 4.1). SERDIF is a combination of tools and processes to enable domain expert researchers to effectively link health and environmental data using a Knowledge Graph (KG) approach (Fig. 4.1). A technology-independent view of SERDIF is presented in Section 4.2.1. Design choices that need to be considered when implementing SERDIF are discussed in Section 4.3. The design choices in implementing SERDIF following a W3C-standard implementation are then summarised in Section 4.4.

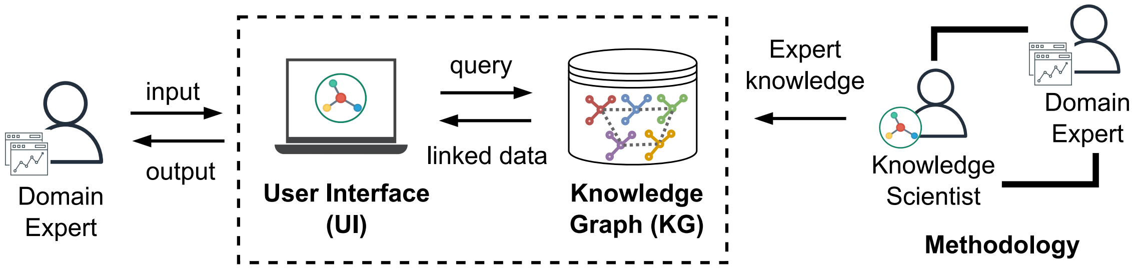


Figure 4.1: Overview of SERDIF and users.

4.1 Expert User Requirements

The initial set of expert user requirements identified in the preliminary study (R1.0, R2.0 and R3.0 in Section 3.3.2) were refined in an iterative manner over three phases.

The implementation and usability testing (and subsequent requirements refinement) undertaken in each phase is detailed in Chapter 5.

The final refined expert user requirements and the motivation for each requirement is described below.

Requirement 1 (R1.2). *Enable HDRs to query environmental data associated with relevant/own individual health events through location and time, within the area of the event and a period of data before the event.*

Researchers acknowledged the complexity associated with integrating relevant environmental data sources and linking them with health events data in the expert meetings and usability testing (*Challenge 1: data interoperability*). The complexity in the integration of environmental data referred to the diverse data input formats and sources, the vast volume of data to ingest in the data pipeline and the generation of a link not based on unique identifiers. The data linkage includes selecting a proper and flexible time window and geographical area to filter the environmental datasets based on the location and date features of particular health events. This process establishes a new link and constructs a new health-environmental KG transforming the raw data.

As discussed in Chapter 3, interacting with a KG requires understanding the underlying graph structure and the complexities associated with the querying language (Section 3.2.1). Even more, when the user needs to write complex queries that use specific elements like location and time to establish new links and construct a KG (*Challenge 2: KG usability*).

Requirement 2 (R2.2). *Support the understanding of event-environmental linked data and metadata, with its use, limitations, and data protection risk for individuals, by using a simplified view focused on the data linkage process with optional further information.*

The linked data generated from a semi-automatic linkage process can be difficult to understand for non-technical users as discussed in Chapter 3. The technologies used to link the data might not provide a transparent pipeline to see each of the steps of the process and the human effect on the linkage. This lack of clarity can lead to the generation of linked data that is not usable for a researcher but also not reusable for other researchers due to lack of transparency of the source of the data and how it has been processed.

Furthermore, the generated linked data may be limited in terms of consistency, may lack a feature of interest or may not be available for an area or time window of study. This data limitation might be unknown if a researcher is only presented with the data, and not with the proper metadata information about the linkage process. In addition, an unclear or uninformed data linkage process can lead to limited evidence to assess the potential data protection risk for individuals, which is especially important for health data. For example, an individual might be re-identifiable given a particular health context where the context has not been provided as metadata.

The role of the metadata in informing the linkage process is key for making the process explainable, traceable, and transparent for researchers. Therefore, researchers would benefit from a simplified and explorable view of the metadata in a human-understandable manner.

Requirement 3 (R3.2). *Export event-environmental linked (meta)data to be used as input in statistical models for data analysis (CSV) and for publication (CSV, RDF).*

Domain experts need to export the data in a usable format for data analysis and publication. During the expert meetings and usability testing, researchers expressed their preference for a data table export to be used as input for their statistical models (e.g. CSV file). Researchers in the health-environment domain are used to working with data tables, which could favour the reuse of the data if published in this format. In addition, the data should also be made available in an interoperable format (e.g. RDF), in step with the current efforts of the open science community to make data Findable, Accessible, Interoperable and Reusable (FAIR) for humans and machines.

4.2 SERDIF – technology independent design

This section describes the Semantic Environmental and Rare Disease data Integration Framework (SERDIF) which was designed to address two key challenges in support of researchers linking health events and environmental data in their workflow. Both *Challenge 1: Data interoperability* and *Challenge 2: KG usability* were reinforced as key challenges through the state of the art analysis as presented in Chapter 3. Essentially this can be summarised as follows: researchers face significant technical challenges when integrating their complex scientific datasets in this domain [Canali and Leonelli, 2022; Ives et al., 2022; Standing Committee on Emerging Science for Environmental Health Decisions et al., 2018]; and while KGs and SW technologies have been used in other domains to address these technical challenges, more research is needed to reduce the expertise required to benefit from KG technologies [Hogan, 2020] (Section 3.4).

As stated in Chapter 1, SERDIF is a combination of three components (Fig. 4.1 and 4.2): Methodology, KG and User Interface (UI). The components have been designed and developed using a User Centred Design (UCD) approach in order to achieve the expert user requirements. The refinement of the requirements, implementation of the individual SERDIF components and usability testing are detailed in Chapter 5.

The SERDIF components are described in this section from **a technology independent perspective** to allow for a wider adoption of the components, by not binding it to a specific set of technologies.

The role of SERDIF within a researcher's health-environment workflow is presented in Fig. 4.2, and will be detailed further in the next subsections.

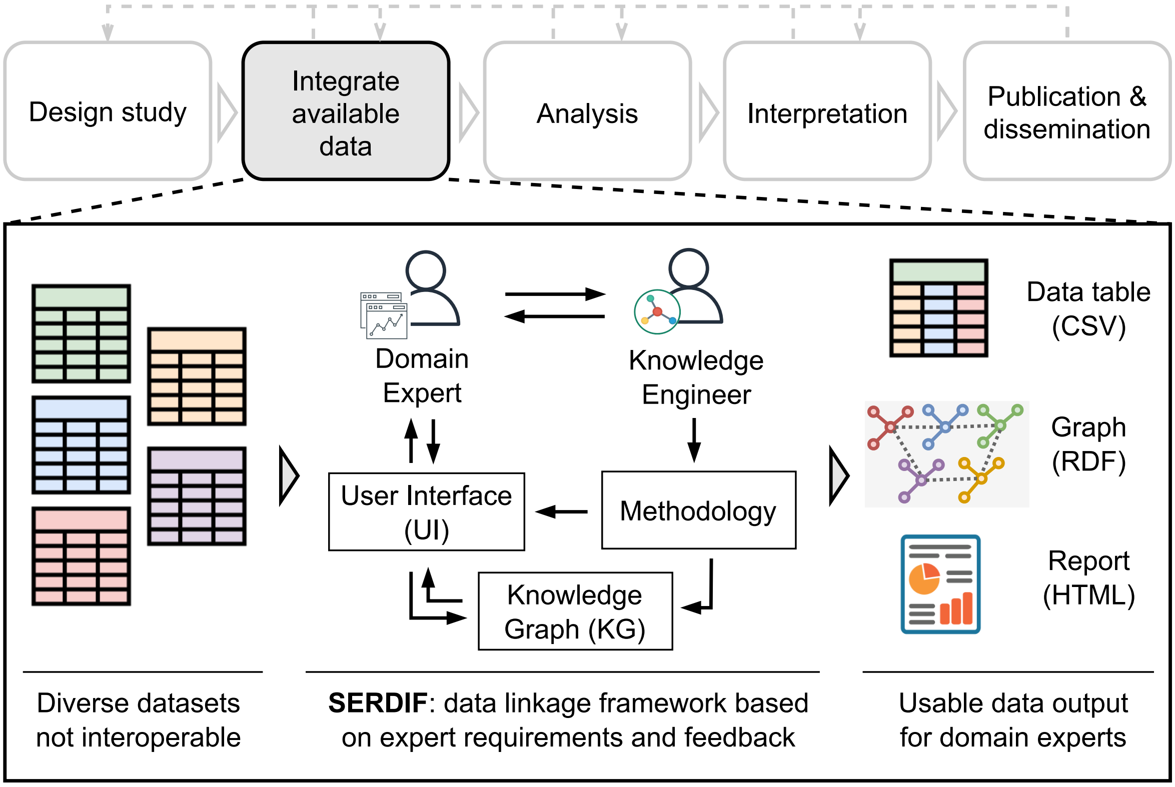


Figure 4.2: Role of SERDIF in researcher's health-environment workflow.

The location of SERDIF within a researcher's health-environment workflow is in the *Integrate available data* process, just after designing the scientific study (Fig. 4.2). The three SERDIF components (Methodology, Knowledge Graph and User Interface) facilitate the process of integrating environmental datasets and linking them with health events for HDRs. The implementation of the framework results in machine- and human-understandable linked data ready for analysis and publication in the following process of the researcher's workflow (Fig. 4.2).

4.2.1 SERDIF Components – Technology independent

The **Methodology** describes a series of steps to facilitate the process of making data interoperable and usable for Health Data Researchers (HDR) (achieving R1.2, R2.2 and R3.2). The methodology includes six sequential steps: (0) preparation, (1) data collection, (2) semantic uplift, (3) data linkage, (4) data interaction and (5) usability evaluation. These steps are described in more detail below. The methodology requires and promotes collaboration between domain experts and knowledge engineers resulting in a usable KG- based approach for domain experts.

The **Knowledge Graph (KG)** component addresses *Challenge 1: Data interoperability* by uplifting the tabular datasets to graphs, importing these graphs to a database and exposing an API for researchers to run queries against the constructed KG (R1.2). In addition, the KG provides the means to explain the linkage process between health and environmental data based on researchers' input, with a view to enhancing the use of the data within appropriate contexts (R2.2). The KG approach influences steps 2, 3 and 4 of the methodology, where a knowledge engineer is needed. How the methodology steps are specialised to the technology binding of W3C standards KG technologies is described in Section 4.3.

The **User Interface (UI)** component is a tool towards addressing *Challenge 2: KG usability* by facilitating the use of the KG component (R1.2). The UI makes the query process intelligible for domain experts (R1.2), the resulting linked data from the query easier to understand given a specific context of use (R2.2), and provides export functionality to retrieve the linked data for analysis and publication (R3.2). The UI tool facilitates the implementation of step 4 of the methodology. Detail on the implementation of the UI of SERDIF over the three iterations of development is provided in the Framework Implementation sections within each usability test phase presented in Chapter 5.

Methodology steps. The methodology steps are presented in this section from a technology-independent perspective without any consideration towards the use of a KG approach.

Step 0: Preparation. Perform the design study phase in an health-environmental workflow, defining the strategy to answer a research question using empirical data. The design study phase also requires a clear definition for the health events relevant to the study, and the permission to process the health event's location and date to link it with environmental data. Another important element is the definition of potential queries to help explore the research question.

Step 1: Data collection. Gather the available environmental datasets relevant to the research question of the study. The datasets are expected to have spatial and temporal features, which are required for *Step 3: Data linkage.*

Step 2: Semantic uplift. Design and execute rules on how to make the environmental datasets gathered in *Step 1: Data collection* interoperable.

Step 3: Data linkage. Define a query template that links the environmental datasets within an area relevant to an event[1] location and selects a period of data before that event date. The query template has placeholders (or variables) for users' input (*Step 4: Data visualisation*) and should be designed to be generic enough to adapt to different data sources.

Step 4: Data interaction. Design an initial User Interface (UI) to allow non-technical users to (i) input the minimum event data required to link with environmental datasets, (ii) specify the user's relevant data linkage variables for the query template defined in *Step 3: Data linkage*, and (iii) execute the data linkage query and export the linked data and metadata generated as a data table for analysis, a graph for publication and an interactive report for exploration.

Step 5: Usability evaluation. Evaluate the usability and potential usefulness of the UI solution defined in *Step 4: Data interaction* in achieving the user requirements. Conduct the evaluation in an iterative manner progressing from version to version until the user requirements are achieved.

4.3 Guidance for implementing SERDIF

For each of the steps in the methodology, guidance is provided to researchers as to what design choices need to be considered when implementing SERDIF for a new use case. This guidance is based on the experience of the author of this thesis during implementation of SERDIF.

4.3.1 Step 0: Preparation

The documents associated with the research's data processing strategy and compliance with data protection laws and regulations need to reflect the use of a KG approach. The data processing information can be typically found scattered across a series of documents based on local, national and international laws and regulations and supervised by data protection authorities. Examples of these documents include Data Management Plans (DMP) and Data Protection Impact Assessments (DPIA). The HELICAL project DPIA template is provided as a concrete example for a data processing document in [Christofidou et al., 2023].

The use of KG in an health-environmental project is summarised below as a general example to include in the relevant data processing documents.

[1] An event is something that occurs in a certain place during a particular time (see Particular health event definition in Section 1.2)

Processing purpose. The processing purpose is to link environmental data with particular health events through location and time for conducting academic research, to better understand the extrinsic factors that influence health outcomes.

Processing summary. The data processing includes linking health events and environmental data using a query defined by the user (Section 4.3.4 and 4.3.5), and exporting the linked (meta)data for analysis and eventual publication as open as possible (Section 4.3.5). The query includes potential personal information as individual location and event date to select the environmental datasets within that area/region, which are then aggregated and filtered for a time previous to the health event. The resulting dataset contains environmental observations associated with a set of health events related to individuals with a particular health context. The dataset export also contains a metadata record describing the information related to the data linkage process regarding the data processing strategy and compliance. For a better understanding of the outcome an example was made available at [Navarro-Gallinad et al., 2023].

Areas of information risk related to the linkage processing. If the health event data is considered personal data by the data controller and/or Data Protection Officer (DPO) assigned to the data, the resulting event-environmental linked dataset is also personal data. The dataset has three personal data elements present: (i) individual dates, (ii) a particular location or region and (iii) the health context; which are enough information to re-identify a person [ICO, 2012]. Effective anonymization of individual-level data is only possible at the expense of the value for research, and even then, people might not be protected against re-identification [de Montjoye et al., 2015].

Compliance requirements. The researchers require explicit permission from the data controller and/or DPO assigned to the health event data to access, use and process the data if the event data is considered personal data. Table 4.1 presents the applicable Information Commissioner's Office (ICO) criteria (i.e. only shown not N/A assessment categories) as the outcome of a DPIA. The Data matching and Tracking criterion were lowered from 'High' to 'Medium' due to the additional measures to mitigate the risks, such as the ones derived for each criterion in Table 4.1. The final assessment selection has to be confirmed by the data controller.

Once the research has been approved by the data controller of the health data, the researcher can advance to the actual data processing, which involves a data integration challenge of multiple and diverse data sources.

4.3.2 Step 1: Data collection

Practitioners are advised to consult a domain expert on the most appropriate environmental datasets to answer the research question identified in the *Design study* process on an health-environmental workflow (Fig. 4.2).

Table 4.1: Applicable 'High Risk' assessment using ICO criteria with the corresponding additional measures to mitigate the risk.

Criterion	**Assessment**	**Comments**
New technologies	Low	Environmental data is linked with personal health data (location and time of an event) using queries on a Knowledge Graph. **Mitigation:** the personal health data is only used to filter relevant datasets.
Data matching	Medium	The location and time from personal health events is used to associate an environmental record to each event. **Mitigation:** the processing is computed on encrypted laptops that access and consult the health data. Event-environmental linked data won't be published as open data, only example data, a generic metadata record together with the workflows and code.
Tracking	Medium	Personal health data used includes the geolocation, date and health context of the event, which are used in the query linkage process. The resulting linked data is considered pseudonymised since effective anonymisation was not possible without losing value of the data for research. **Mitigation:** same as Data matching mitigation.

Based on the experience of the author of this thesis, the following metadata needs to be gathered and fully documented: the information related to the dataset descriptors (e.g. licence, title, version, temporal and spatial information and structure of the dataset), data provenance (e.g. distribution and download url), data use, agents that downloaded the data (e.g. researcher, software and entity) and the definitions for the environmental variables, including the units and source of the information. The relevant metadata content depends on each use case and additional information might be needed for particular cases. The metadata selection criteria from the author of this thesis should be taken as the basis to build on.

In addition, the geometries for the relevant areas of study are recommended to be gathered if the spatial linkage, defined in the *Design study* process (Fig. 4.2), requires the integration of environmental datasets for specific areas (e.g. counties, provinces, prefectures or countries).

If suitable datasets are available as RDF graphs the recommendation is to reuse them, instead of downloading the tabular or compressed distribution of the dataset. Practitioners can then skip the first three substeps in *Step 2: Semantic uplift* and resume from *Host the RDF (meta)data.*

4.3.3 Step 2: Semantic uplift

The author of this thesis recommends to uplift the environmental datasets to RDF, the W3C standard for data interchange. The design choice of using RDF against other interoperable approaches relies on being an open standard considered as best practice for data exchange and publication in the Web, increasing the flexibility and scalability of the graph approach. As introduced in Chapter 1, RDF can use RDFS and OWL languages to describe groups of related resources, relationships between these resources, and provide meaning (semantics) and logic rules understandable by machines. These two languages built on RDF are also part of the W3C standards that materialise shared semantics across the Web. Furthermore, this design choice promotes open science, which is gaining importance to achieve shared goals across scientific disciplines. This design choice for using W3C standards is described in detail in Section 3.1.

The uplift process can be divided into three subprocesses: (1) define the semantic (meta)data model, (3) uplift the (meta)data and (4) host the RDF (meta)data.

Define the semantic (meta)data model. A set of vocabularies and ontologies need to be identified to represent the (meta)data for the domain with the objective of answering the research question from the study. Environmental datasets can be described using the Semantic Sensor Network (SSN) ontology [Haller et al., 2017], for describing sensor observation data, or the RDF Data Cube vocabulary (QB) [Cyganiak and Reynolds, 2014], for describing statistical and multi-dimensional datasets; both W3C standards. The author of this thesis design choice is to use the QB vocabulary due to being a more general and flexible semantic model. QB includes enough flexibility to represent different types of environmental exposures from sensor data, occupational exposures, self-reported outcomes, diet, built environments and others, in the shape of surveys or spreadsheets.

The datasets can be represented using QB as a collection of observations through repeated measurements over time (i.e. time series) associated with a fixed location [Cyganiak and Reynolds, 2014]. The environmental variables can be represented as a *qb:MeasureProperty*, including the necessary comments, description, units and URLs to the original source of information or standard vocabulary that describes the specific environmental variables. Each multi-measure time series can be represented as a *qb:Slice* using QB with a location defined as a Geographic Query Language for RDF Data (GeoSPARQL) geometry [Perry et al., 2012]. GeoSPARQL is the standard for representation and querying of geospatial linked data for the Semantic Web from the Open Geospatial Consortium (OGC). GeosPARQL enables the spatial reasoning needed in the recommendation for *Step 3: Data linkage*. Therefore, the geometry data needs to be represented as GeoSPARQL geometries as well. In addition, the time dimension of the collection of observations can be represented using the *xsd:dateTime* data type, enabling the temporal reasoning, which is key to answer HDR research questions.

Regarding the metadata associated with the environmental datasets, the author of this thesis recommends the use of the Data Catalog Vocabulary (DCAT) [Albertoni et al., 2020] for the dataset descriptors, the PROV Ontology (PROV-O) [Lebo et al., 2013] for the dataset provenance and agent that downloaded the (meta)data, the Open Digital Rights Language (ODRL) for the data use [Iannella and Villata, 2018]. The three vocabularies and ontologies are W3C standards.

The choice of vocabularies and ontologies to describe the environmental dataset and metadata depends on the (meta)data gathered on *Step 1: Data collection*, further W3C standards are available here: `https://www.w3.org/TR/?tag=data&status=REC&version=latest`. Practitioners should first check if an existing semantic model can be reused before generating a new one. Generating a semantic model can be complex and time-consuming as it requires domain experts and knowledge engineers to collaborate in an efficient manner [Jacobsen et al., 2020b].

Uplift the data. Given the environmental domain, most of the datasets gathered in *Step 1: Data collection* will probably be stored as tabular, relational, or compressed data depending on the purpose, intended end-user and volume of the datasets. Tabular formats include simple data tables stored as TSV or CSV files where every record shares the same set of variables, easily readable for humans as information is stored in plain text. Relational data can store multiple data tables in a Relational DataBase (RDB) without a particular internal structure (i.e. data tables can have different variables) when the data volume is too large to be stored in memory. The RDBs allow humans to interact with the data by running Standard English Query Language (SQL) queries. Compressed files such as NetCDF and GRIB were created for transmitting large volumes of (gridded) data between computer systems with an efficient storage and retrieval format for autonomous systems. However, these compressed files present a higher technical barrier for humans than the previous two data formats.

Two W3C standards cover the uplift of tabular and relational data to RDF, direct mapping [Arenas et al., 2012] and RDB to RDF Mapping Language (R2RML) [Das et al., 2012]. Direct mappings define a set of simple transformation rules that can be used to generate RDF graphs from relation data. Given a RDF triple structure of subject–predicate–object (Section 1.1) and a data table, a direct mapping transforms a row in a RDF subject or object, a column name in a RDF predicate and a cell in a value. This transformation does not allow for customisation as the rows, columns and cells will have the same names and values as in the RDB, with a ready-made structure and vocabulary. The ready-made structure creates problems in terms of reducing the complexity for the RDF structure, which may be needed to represent a particular concept; and the vocabulary is only known to the agent that performed the direct mapping process, complicating semantic interoperability. R2RML was developed to address these issues by allowing the annotation of RDBs with existing vocabularies and ontologies adding customisation to the structure and vocabulary. R2RML specified an ontology to define the mapping files and an interpretation of the mappings to generate RDF files. The R2RML mappings are themselves expressed as RDF graphs.

A couple of non-W3C standard alternatives are worth mentioning in the case they can be useful in certain situations or if they become a standard after the publication of this thesis. The RDF Mapping Language (RML) [De Meester et al., 2022] is an extension of R2RML with the aim to extend the applicability, scope and support input data in other formats beyond relational databases. These data formats include CSV, TSV, XML and JSON. RML is a draft of a potential specification (i.e. future W3C recommendation) as of the submission of this thesis. Another path to uplift the data would be to use a programming language such as Python or R, where a data uplift script could be used to use any non-RDF data source to RDF. The uplift script can read the input data and convert it into RDF by using templates or logic. However, the resulting RDF file would need to be validated in terms of syntax and against the vocabularies and ontologies constraints. The syntax validation can be done by using a RDF validator tool such as [IDLab, 2017; Joinup, 2023; Prud'hommeaux, 2006] or an available library from the programming language. The validation against a vocabulary and ontology can be done by using the SHApes Constraint Language (SHACL) [Knublauch and Kontokostas, 2017], which is the W3C standard language for validating RDF graphs against a set of conditions for nodes and properties. For example the SHACL shape can check if a value node is of a given data type or within the adequate range of values, which could be useful if the raw data has not been validated by a certified entity.

The author of this thesis design choice is to use R2RML to uplift the non-RDF data to RDF graphs granting the possibility of using existing vocabularies in the *Define the semantic (meta)data model* subsection in this step. Furthermore, the R2RML mapping is a declarative approach that provides the practitioners with the mapping document to serve as provenance and that can be annotated with metadata or queried for data quality checks, whereas a using programming language would not.

While R2RML was designed for relational databases, practitioners can use the R2RML-F engine [Debruyne and O'Sullivan, 2016] that allows access to CSV files and relational databases. Furthermore, the R2RML-F this engine has a functionality to transform the raw data from the CSV files within the same mapping file, providing traceability for the pre-processing of the raw data. For example, raw datetime values can be transformed to a valid standard syntax for RDF files (2022-10-01 12:00:00 to 2022-05-20T12:00:00Z). This approach was evaluated by Crotti Junior et al. 2017 against state-of-the-art CSV uplift tools yielding positive outcome [Crotti Junior et al., 2017a]. However, an extra pre-processing step is needed to convert the environmental compressed formats (e.g. NetCDF and GRIB) to CSV to be able to use R2RML.

A sample of the R2RML mapping to uplift an air pollution dataset from CSV to RDF is made available in Appendix A and in the thesis' GitHub. The RDF files resulting from running the mapping have been published as open data in Zenodo [Navarro-Gallinad, 2021].

Host the RDF (meta)data. The resulting uplifted graph files are recommended to be stored in a database to enhance the efficiency of querying large datasets, which is the case for environmental data. Based on the best practices for data access on the Web from W3C [Farias Lóscio et al., 2017], data and metadata as RDF graphs are recommended to be hosted in a triplestore that provides a SPARQL query editor and a RESTful API, both with the proper documentation to facilitate the use. The query editor and API should allow the export of a bulk datasets or subsets in different machine- and human-readable formats. The triplestore service should be made available on the Web in real-time, specifying the update frequency and if some data is not available. In addition, the triplestore of choice needs GeoSPARQL support to enable the queries from the next step (*Step 3: Data linkage*).

Hosting the graph data on a database can link different graphs in an automatic manner if these are described using the same vocabularies and ontologies, or if an ontology that establishes relationships between these ontologies has been also imported to the triplestore. In this research, environmental data from different sources are described using the same ontologies as described in the previous subprocess. This feature can result in linking the data with simpler queries as a link has already been established between graphs. In a similar manner, when reusing ontologies, graphs can establish links with other endpoints as the machines will understand the resources belong to the same group, which is the case for interlinking.

Step 2: Semantic uplift is considered the key step to address *Challenge 1: Data Interoperability* and the most complex step of the SERDIF methodology in terms of knowledge and technical expertise required to define the semantic (meta)data model and the mappings to answer particular queries relevant to a given context.

The technical complexity includes (i) the variety of data sources to harmonise to the chosen semantic (meta)data model, which usually means that an uplift mapping needs to be defined for each data source as combining declarative and automated approaches is challenging [Chaves-Fraga and Dimou, 2022]; (ii) the consistent reuse of vocabularies in the mapping, which often contains large numbers of restrictions on the usage of the classes and properties [Dimou et al., 2015; Heyvaert et al., 2019]; (iii) the steep learning curve associated to developing mappings, which can be syntactically heavy and not intuitive leading to errors in the process [Crotti Junior et al., 2017a,1; Randles and O'Sullivan, 2022]; (iv) the mapping quality, which is often the root cause of data quality issues as it is a complex task prone to errors [Crotti Junior et al., 2017a; Randles and O'Sullivan, 2022]; (v) the often large volume of environmental data (e.g.

weather and air pollution); (vi) data sources can be dynamic depending on the use case, which can generate misalignment between the mappings and input data. In addition, this step requires the collaboration of domain experts and knowledge engineers to effectively uplift the non-RDF data to RDF. The strategies adopted by the author of this thesis to address the technical complexities associated with *Step 2: Semantic uplift* are presented in Table 4.2, which are grouped based on the phases of the mapping process [Debruyne et al., 2015].

Table 4.2: Summary of the technical complexities associated with the semantic uplift process annotated with strategies to address the complexities and grouped based on phases of the mapping cycle.

Phases of the mapping lifecycle	**Technical complexities of the Semantic uplift process**	**Strategies proven to work**
Stage	Knowledge engineer and domain expert collaboration	Host project meetings where both parties are present where the semantic uplift process can be discussed as it progresses and expert requirements refined.
Characterise	Variety of data sources to harmonise	Examine and understand the data sources with respect to what each data entity represents and characterise the difficulties in finding matches and alignments between the data models. Again requires both workshops with Domain Experts and Knowledge Engineers.
Reuse	Variety of data sources to harmonise	Reuse several standardised vocabularies as recommended by the W3C data on the web best practices [Farias Lóscio et al., 2017] and involve domain experts to facilitate the correct semantic reuse of existing vocabularies.

Match	Deciding if matching algorithms will be helpful to suggest possible correspondences between complex datasets	This step will also be influenced by discussions during the Characterise step and whether using software matchers to propose possible matches between elements of the data models will be helpful or not. Otherwise a manual process involving Knowledge Engineer and Domain Expert is needed to find the data model correspondences and document them.
Align and map	Developing mappings steep learning curve	Use visual existing tools like Juma [Crotti Junior et al., 2017b] and RML editor [Heyvaert et al., 2016] to guide the process of developing mappings based on correspondences documented from the *Match* step.
	Mapping quality	Validate the mapping file with the Mapping Quality Framework [Randles and O'Sullivan, 2022] and check a small representative RDF graph resulting from the refined mapping against SHACL constraints for the schema (RDFS), vocabularies and ontologies (OWL). Dataset quality assessment approaches such as RDFUnit [Kontokostas et al., 2014] and Luzzu [Debattista et al., 2016] can help in the identification of quality issues from the source data or mapping.
	Large datasets	Subset large datasets into a manageable size during data collection or uplift process with a temporary dataset, and define a mapping template with placeholders for the temporary datasets generated.

Application	Alignment of mappings during an ongoing project	Keep in sync the mapping with the underlying data sources [Dimou et al., 2016], identify alignment issues and determine when the mapping requires updating or re-execution with the Mapping Quality Framework [Randles and O'Sullivan, 2022].

4.3.4 Step 3: Data linkage

Given the design choice of uplifting the datasets to RDF in *Step 2: Semantic uplift*, the query template is recommended to be written in the SPARQL Protocol and RDF Query Language (SPARQL) [W3C SPARQL, 2013], the W3C standard for querying RDF files. A SPARQL query can integrate the environmental datasets and link them with the health events through their common elements, location and time. As described in Chapter 2, A SPARQL query template can be prepared to infer relationships between the health events and environmental data based on location and time. For example, environmental datasets can be linked spatially, based on the location of the event by computing the shortest distance or within a geographical/administrative area that contains the event, and temporally, based on a time window related to the event date (Listing 4.1). The spatial inference can be conducted by the use of GeoSPARQL functions such as *geof:sfWithin*, which select the geometry (e.g. lat/lon point) that is included in a larger geometry (e.g. geographical area) (Listing 4.1). The temporal inference from the *xsd:dateTime* data format can be used to define time windows from a particular date by subtracting a temporal duration as in Listing 4.1, or to extract the month or year for a later aggregation [W3C SPARQL, 2013].

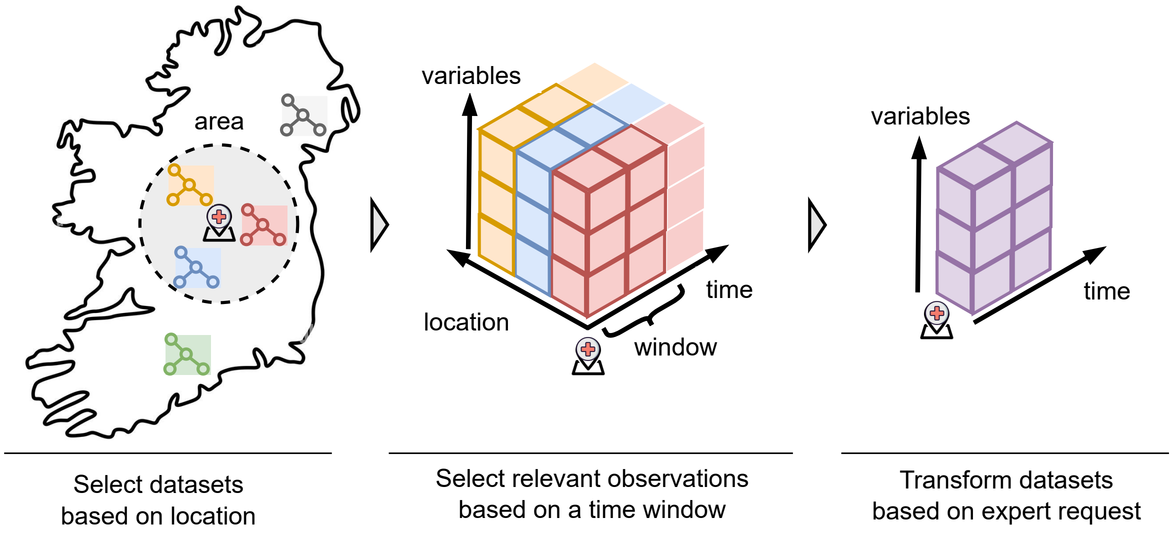
Other inferences can be computed at the SPARQL query level from known patterns based on the current progress of the research and knowledge from the domain experts, advancing towards an expert system. Examples in the health-environmental domain include checking for a data point over a certain threshold that may be considered harmful or if a number of variables is present in a certain combination that has proven to lead to specific health outcomes for an event (e.g. high risk of relapsing). This inferences can be computed by defining an ad-hoc ontology specific to the use case or using SHACL to define shapes in a more flexible approach. Furthermore, SHACL can also be used in SPARQL queries to harmonise the uplifted datasets in the case that they have not been uplifted from the same source. In the case presented in this section, the datasets have been uplifted using the same vocabularies and ontologies to a common structure and they were not available as RDF from the beginning. However, if the requirement is to link them further with existing datasets in other endpoints (i.e. interlinking), these would have to be harmonised at the SPARQL query level, which can be done with SPARQL query that includes some logic statements, SHACL or by defining a mapping between the ontologies.

Listing 4.1: Spatiotemporal reasoning used in the SERDIF querying process as a SPARQL example.

```
# -- Namespaces ----------------------------------------
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)
```

Once the datasets are linked, the set of environmental datasets linked with each health event can be integrated by using the SPARQL aggregate functions (e.g. SUM, AVG or MAX) [W3C SPARQL, 2013] as represented in Fig. 4.3. This process transforms the raw data into an aggregate set of environmental variables for each health event, which was reported to be a usable structure for domain experts (Chapter 5). The data transformation process within the SPARQL query is part of the definition of enable the data linkage for HDRs from Section 1.2 (highlighted below).

Enable the data linkage: data linkage is defined as bringing together from two or more different sources, data that relate to the same individual, family, place or event [Holman et al., 2008]. In this thesis enabling data linkage refers to providing the means for a HDR to (a) make their datasets interoperable as RDF files, (b) create queries to link datasets to a particular event based on the spatial and temporal aspects of the data and (c) **export a transformed view of the linked data in a usable format for humans and machines**.



Select datasets based on location

Select relevant observations based on a time window

Transform datasets based on expert request

Figure 4.3: Simplified view of *Step 3: Data linkage* for a single health event.

However, the type and need of data transformation depends on each use case and should be discussed with the experts. Therefore, knowledge engineers are recommended to co-design the SPARQL queries with domain experts towards matching the structure and format with the expert/system requirements.

In addition to the main query, further SPARQL queries can be executed to compute the historical mean values of the environmental variables by using a query with aggregate algebra functions (e.g. count, sum or average).

The *Construct* query form is recommended as it returns a single RDF graph specified by the graph pattern of the query, it constructs a new KG. The *Construct* SPARQL query can represent the event-environmental data and metadata using the same vocabularies and ontologies from *Step 2: Semantic Uplift*. Given that the target is health related data, information described with ODRL vocabulary can be extended with the Data Use Ontology (DUO) [Courtot, 2021] to specify the data use, and the Data Protection Vocabulary (DPV) [Pandit, 2022] to describe the use and processing of personal data, if the linked health data is considered personal data.

Furthermore, the *Construct* query form provides the flexibility for researchers to describe the resulting linked data with other vocabularies and ontologies. Researchers seeking to exchange the data in the healthcare sector are encouraged to describe the health data compliant with Health Level Seven (HL7) and Fast Health Interoperability Resources (FHIR) standards for healthcare data exchange [HL7, 2022]. The use of the HL7 FHIR standard can provide the linked data with connections to medical terminologies defined in existing ontologies ready to be shared with the healthcare sector besides HDRs.

The design choice for this step emphasises the authoritative component of linking the data at the query level while using ontologies to link the data, defined in *Step 2:*

Semantic Uplift. This choice grants flexibility for ongoing research projects, where a basic set of links is established between datasets described with the same ontology; but new links, that were not thought of from the start, can be explored by constructing a new KG. This means more complex queries and less performance as a new KG needs to be constructed, but it grants more flexibility. Therefore, the approach taken depends on the use case, based on the progress made in the research project it might be worth combining different approaches.

4.3.5 Step 4: Data interaction

The author of this thesis recommends designing a simplified graphical UI that facilitates the process of populating the SPARQL query template with expert inputs, while hiding the technical complexities associated with the data linkage process. The design choice for the UI is supported by the positive usability results against the expert requirements reported in Chapter 5 and does not need to be based on W3C standards.

Health event input or upload (R1.2). The health events relevant to experts' research can be input or uploaded to link them with environmental datasets. The minimum information needed includes the event identifier, geographical coordinates (latitude and longitude), date of the event, and the time window parameters to gather environmental data relative to the date of the event (length, lag).

Data linkage options (R1.2). A series of data linkage options should become available for the user to specify the metadata associated with the linkage process and the statistical methods to link the environmental data to the health events. The minimum options for the data linkage process metadata include the dataset descriptors, dataset use, dataset provenance and processing of the data to generate the linked data using the ontologies and vocabularies described in *Step 2: Semantic uplift.* Regarding the methods to integrate and subset the environmental data based on the health events input, the recommendation is to add the necessary spatial and temporal options requested in each use case including the aggregation methods if two or more environmental datasets are relevant for a single event.

Export output (R2.2, R3.2). The recommendation for the export output is that the linked data and metadata is exported in a (re)usable manner for the domain experts. A collection of three document types should be made available as an export for the researchers: (i) the linked data ready for analysis as a data table (CSV) and graph (RDF), (ii) the metadata describing the linkage process and the linked data for transparency and informed data use (CSV and RDF), and (iii) the interactive report to make the linked data and metadata easier to understand before starting the analysis, including a section to explore and visualise the data within environmental context (HTML). The environmental context should provide information about the magnitude of the value compared to the historical average for the region of the event and time period. A complete export following these recommendations is made available at [Navarro-Gallinad et al., 2023].

The collection of datasets and documents exported as a result of formulating a query follows the W3C best practices form data on the Web [Farias Lóscio et al., 2017]. Table 4.3 summarises the evidence towards following the W3C best practices.

Table 4.3: Summary of how the data on the Web best practices from W3C are met in the SERDIF UI data export.

W3C Best Practice (#)	**Evidence**
Metadata (1-3)	Metadata described reusing existing standard terms and vocabularies, and provided as machine-readable (RDF) and human-readable (CSV and HTML Web page) formats. The metadata includes dataset descriptors (DCAT) and structure of a distribution (QB).
Data licences (4)	Licence information is present to assess the data use (DCT) including the rights to use the data (ODRL).
Data provenance (5)	Provenance information of the data linkage and the data processing steps is available (PROV-O).
Data quality (6)	Domain experts assessed the suitability of the dataset to be used in their research workflows.
Data versioning (7,8)	Versioning data included as part of the metadata (DCT) and included in the dataset URI. A comment describing the version of the dataset is also provided.
Data identifiers (9-11)	Datasets are identified with a URI that includes the type of dataset and version date query time readable for humans. A placeholder is included within the dataset metadata for an external persistent URI in case the data can be published in a data repository.
Data formats (12-14)	Datasets are available as machine-readable (RDF) and human readable (CSV and HTML Web page) formats. Locale-neutral data structures and values (XSD datatypes) are used to represent the dataset observations. Contextual information about the dataset observations (historical averages) is provided to promote the reuse in different contexts.

Data vocabularies (15, 16)	Standard vocabularies are used to describe the data and metadata (QB, GeoSPARQL, PROV-O, DCAT, DCT, ODRL).
Data access (17-27)	Data access practices depend on the personal aspects of the data and the permissions to publish the data, granting data processing permission to a third party. A dataset accompanied with metadata, resulting from the linkage process, is made available in a data repository provided with a DOI as an example [Navarro-Gallinad et al., 2023].
Data preservation (27, 28)	The datasets generated after the query process are provided with a placeholder for an external persistent URI in case of publication in a data repository.
Feedback (29, 30)	The evaluation results with comments about the usability of the datasets are included in Chapter 5 and published in peer-reviewed workshops, conferences and journals [Navarro-Gallinad et al., 2020,2,2].
Data enrichment (31, 32)	An interactive report displays the information in the dataset as visualisations, tables and summaries, in a human-readable format (HTML Web page)
Republication (33-35)	The original publishers and datasets are cited in the data linkage process activity, and their licensing requirements followed for appropriate data use.

Command Line Interfaces (CLI) and Menu-Driven Interfaces (MDI) are two types of user interfaces worth exploring as they might be more usable given a particular context for a certain group of researchers. CLI can be potentially useful when performance is the key aspect of usability and the users are experienced in the command language. Examples include technical researchers, like data analysts, that have already become familiar with the linkage process after sending multiple queries and would prefer to improve the performance of running the query programmatically directly to the API of the triplestore. MDI could be useful for more novice researchers as it provides self-explanatory menus and removes the need to remember the list of manual commands, as in a CLI. However, CLI and MDI can limit the functionalities available to the users involved in a requirement, which reduces the usability of the solution for the experts.

4.3.6 Step 5: Usability evaluation

The best practices in usability testing recommend the combination of summative and formative conceptualizations of usability within an iterative design process centred in achieving the expert user requirements [Dadzie and Rowe, 2011]. By including both conceptualisations the resulting KG approach and UI tool can be evaluated in terms of the effectiveness, efficiency, satisfaction, detecting usability problems and design interventions to minimise them. Furthermore, standard usability metrics such as post usability questionnaires are recommended to track the progress of the satisfaction aspect across each iteration of the study [Sauro and Lewis, 2016].

Domain experts can complete a series of tasks designed with real workflows in mind to evaluate the usability of the framework, which should be written as simple and concise as possible. Even if the domain or nature of the problem is complex, simple instructions and design are more appealing to users. The domain experts need to be included from the start of the user-centred approach (Chapter 5). If domain experts can relate their research and benefit from the KG approach and UI, they will be more keen to invest their time providing more relevant feedback during the evaluation step (*Step 4: Data interaction*).

Other viable choices for an evaluation could have been to use competency questions and performance evaluations. Competency questions are user-oriented questions to scope the ontology or knowledge base that would answer the questions through exploration and queries. Performance evaluations focus on how to optimise a system to provide a reliable and useful outcome aligned with an strategic goal.

In this usability evaluation, domain experts can raise usability problems and performance issues while completing a series of real tasks. This approach provides an evaluation for user requirements and an acceptable performance with a qualitative threshold, instead of a quantitative estimate. Further recommendations on conducting this step are described in Chapter 5.

4.4 Chapter Summary

The design choices made for the SERDIF implementation are summarised in Table 4.4. The design choices listed are: Knowledge Graph approach based on W3C standards (KG-W3C), Generally Accepted Methods (GAM) and Thesis Evaluation Results (TER). TER refers to choices based on iterative results at the end of an evaluation phase presented in Chapter 5 of this thesis.

For this thesis a technology specific approach, W3C-KG, has been selected for implementing the SERDIF KG component. The choice of using a graph data model grants more flexibility to model the domain as the focus is on generating relationships between data points that can generate new insights from the existing data. This flexibility enables more complex queries to be used for scientific research. The science community is moving towards interdisciplinary approaches to answer research questions that require the integration of data sources generated with different purposes in mind, increasing the linkage complexity as in health-environmental. The choice of following a W3C technology implementation makes the graph data model understandable for machines with shared semantics and standards. Machines are provided with enough context to effectively interpret the intended meaning of the data towards higher levels of interoperability (Section 3.1). This poses an advantage in a discipline with vast amounts of data available such as health-environmental where machines can effectively perform processing tasks that a human would not be able to do.

Table 4.4: Summary of the SERDIF methodology implementation, annotated with design choices: Knowledge Graph approach based on W3C standards (KG-W3C), Generally Accepted Methods (GAM) and Thesis Evaluation Results (TER).

SERDIF methodology step (design choice)	**W3C KG-based Implementation**
Step 0: Preparation (KG-W3C)	□ Include the KG approach as part of the data processing strategy and compliance
Step 1: Data collection (GAM)	□ Consult domain expert □ Include dataset metadata □ Download geometry data for relevant study areas

Step 2: Semantic uplift (KG-W3C)	□ Uplift data to RDF [Cyganiak et al., 2014] □ Include QB [Cyganiak and Reynolds, 2014], GeoSPARQL [Perry et al., 2012], PROV-O [Lebo et al., 2013], DCAT [Albertoni et al., 2020], DCT [DCMI, 2020] and ODRL [Iannella and Villata, 2018] for the semantic (meta)data model. □ Define uplift mapping using R2RML mapping language [Das et al., 2012] □ Execute the mapping with R2RML-F [Debruyne and O'Sullivan, 2016] □ Store RDF files in a triplestore with GeoSPARQL support
Step 3: Data linkage (KG-W3C)	□ Link data through location and time using a SPARQL [W3C SPARQL, 2013] query template with GeoSPARQL and xsd:dateTime [Peterson et al., 2012] reasoning □ Return results from a *Construct* query form □ Additional queries for environmental context
Step 4: Data interaction (TER)	□ Design a simple UI on top of the KG focused on the data linkage process □ Include data table (CSV), graph (RDF) and interactive report (HTML) outputs of the linked data and metadata.
Step 5: Usability evaluation (GAM)	□ Combine summative and formative conceptualizations of usability □ Focus on achieving the expert requirements □ Use of standard usability metrics

Chapter 5

SERDIF Evaluation and Implementation

This chapter describes the findings of the usability study, consisting of three usability tests, undertaken to evaluate the usability and potential usefulness of SERDIF to support health data researchers (HDRs) link health events and environmental data. The implementation of SERDIF undertaken for each phase is described within each usability test section.

Usability: *Extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.*
Effectiveness: *accuracy and completeness with which users achieve specified goals.*
Efficiency: *resources used in relation to the results achieved.*
Satisfaction: *extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations.*

ISO 9241-11:2018

Usefulness: *degree to which a user is satisfied with perceived achievement of pragmatic goals, including the results of use and the consequences of use.*

ISO/IEC 25010:2011

It is important to note that the usability aspects and potential usefulness of SERDIF were evaluated for a specific group of health data researchers (*users*) to conduct data linkage tasks (*goal*) towards studying the environmental factors associated with health events related to rare diseases (*context of use*).

The usability testing included **static evaluation elements** and **dynamic evaluation elements**. Static elements are defined in this research as the elements that remain the same for each usability test throughout the phases of the usability study.

Static evaluation elements included the strategy and goals to achieve in each of the usability tests (*Strategy*), the type of participants that would find SERDIF most useful (*Participants*), the description of the experimental setup to conduct the sessions (*Experimental Setup*), the metrics gathered during the experimental session (*Metrics*), and the data analysis methods and instruments (*Data Analysis*) to evaluate the usefulness of SERDIF based on the expert participants (Section 5.1).

Dynamic evaluation elements were progressively incorporated in each phase of the usability study. The dynamic evaluation elements include the summary description of an health-environmental use case that would benefit from using SERDIF (*Use Cases*), the number of participants in each usability test (*Sample size*), the updates on the experimental methodology after each usability test (*Experimental Setup Update*), and the tasks given to the participants to complete using SERDIF (*Participant Tasks*).

The three usability tests (Section 5.2, 5.3 and 5.4) are described following the four steps of a User-Centred Design (UCD) [Jokela et al., 2003] as per (Fig. 5.1):

(i) a description of how the users may use SERDIF for their research (*Context of use*, Section 5.2.1, 5.3.1 and 5.4.1);

(ii) the identification and refinement of the expert user requirements (*Expert user requirements*, Section 5.2.2, 5.3.2 and 5.4.2);

(iii) the implementation and refinement of the SERDIF components based on the requirements (*Framework implementation*, Section 5.2.3, 5.3.3 and 5.4.3);

(iv) the usability evaluation of SERDIF for the use case described (*Evaluation against requirements*, Section 5.2.4, 5.3.4 and 5.4.4). The evaluation step is divided in three substeps including the statement of the quantitative and qualitative results of the usability testing (*Results*); the resulting themes and findings of the thematic analysis (*Thematic analysis*); and the evaluation conclusion on the progress made towards matching the users' context of use and the achievement of the requirements (*Conclusion*).

The details of the usability study was submitted to and gained approval from the School of Computer Science and Statistics (SCSS) Ethics Committee in Trinity College Dublin (Appendix B).

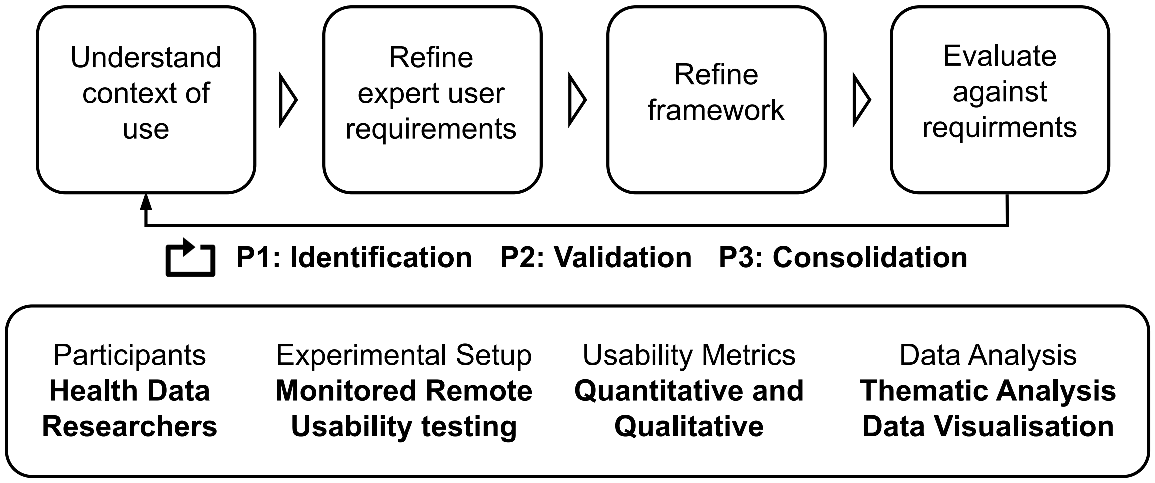

Understand context of use
Refine expert user requirements
Refine framework
Evaluate against requirments
P1: Identification P2: Validation P3: Consolidation
Participants
Health Data Researchers
Experimental Setup
Monitored Remote Usability testing
Usability Metrics
Quantitative and Qualitative
Data Analysis
Thematic Analysis
Data Visualisation

Figure 5.1: Overview of the evaluation approach following a user-centred design. The replay symbol represents the iterative phases Phase 1 (P1), Phase 2 (P2) and Phase 3 (P3).

The remainder of this chapter describes the design limitations of the SERDIF data linkage approach (Section 5.5) and summarises the findings of the usability evaluation (Section 5.6).

5.1 Static Evaluation Elements

The static evaluation elements shared across the three usability tests are described in this section.

5.1.1 Strategy

The objective of the usability testing approach was to evaluate the usability and potential usefulness of SERDIF in addressing the *Challenge 1: data interoperability* and *Challenge 2: KG usability* for ongoing rare disease research involved in health-environmental studies (Section 1.1). The usability testing approach consisted of three phases (Fig. 5.1): (Phase 1, P1) identifying and refining the initial user requirements and framework, (Phase 2, P2) validation of the usefulness of the framework for HDRs and (Phase 3, P3) consolidation of the requirements and framework as a solution for HDRs. The approach combines summative and formative conceptualizations of usability within an iterative design process, following best practices [Lewis, 2014].

Summative usability. The goal is to evaluate the framework with performance and satisfaction usability metrics. The metrics are associated with the completion of a series of tasks by representative users (Section 5.1.2) to evaluate the usability in terms of effectiveness, efficiency and satisfaction. This conceptualisation grants the study a starting benchmark to be compared in each iteration. The quantitative usability metrics included in the study are presented in Section 5.1.4.

Formative usability. The goal is to detect usability problems and design interventions to reduce these problems. The study advanced in a progressive manner through each phase, including researchers and the use case from the previous phase. The iterative approach improved the chances to find errors, ambiguous information and confusing features while generalising the health data input capabilities of the framework. The cultural effect on the usability results was reduced by progressively increasing the sample size and research groups involved. Progressively throughout the phases the coverage of the environmental data was also increased from a single country (Ireland) to multiple countries within Europe, according to the country of interest in each use case. The qualitative usability metrics included in the study are presented in Section 5.1.4.

In the iterative process, the supervisors of this thesis and colleagues with experience in developing full-stack web applications and conducting usability studies tested the tasks and framework as expert usability evaluators. At least three experts usability evaluators were included during the testing stage. Furthermore, a stopping rule was included to prevent infinite iterations. The usability study in this thesis was considered completed when the expert requirements did not require any further refinement (i.e. consolidated requirements) and were achieved. The evidence for the stopping rule combined quantitative and qualitative usability metrics, providing a reasonable expectation that undiscovered problems will not lead to drastic consequences in the future. The study also took into account the time constraints associated with the PhD and the feasibility of the expert requests in terms of current technology limitations.

5.1.2 Participants

The participants were Health Data Researchers (HDRs), meaning researchers with a health background and statistical or data analysis experience (e.g. clinicians, health informatics technicians/managers and epidemiologists), or statisticians and data analysts that are studying health related outcomes. In addition, only HDRs without practical expertise in using Knowledge Graphs (KG) were eligible for the study.

A list of eligible researchers was drawn up based on those involved in the target use case projects (see each of the *Context of Use* subsections for details). The eligible researchers were recruited by sending two personal emails asking if they would like to participate in the usability study. The first email contained a brief introduction and the rationale of the experiment, and expert participants were asked to complete an online poll to select a time slot that worked for them. The second email provided the participants with the details of the experiment: virtual meeting details, participant information sheet and consent form to be read and signed, the User Interface (UI) link and login credentials.

5.1.3 Experimental setup

The expert participants that agreed to participate were asked to sign the consent form at the start of the virtual meeting if they had not already returned it signed by email. The usability sessions were conducted remotely using a video conferencing platform where participants were asked to share their screen and audio while observed and assisted, if necessary, by a usability moderator (the author of this thesis). The participants were asked to complete a series of tasks during the sessions. The tasks were gathered and curated before the experiment was conducted and were derived from consensus among HDRs with real workflows in mind. The tasks included the processes to link the environmental and health data for research based on the expert user requirements.

The participants were briefed about the experimental protocol, which included two evaluation instruments: a think-aloud protocol and post usability test survey.

Think-aloud protocol. The think-aloud protocol consists of completing the tasks using the SERDIF UI while thinking aloud [Boren and Ramey, 2000]. The think-aloud protocol was new to the participants and it was explained in simple words as a way to communicate their thoughts as they progress through the UI completing the tasks. The usability moderator made clear that the participants should not seek help from the moderator nor ask direct questions but if they became blocked in progressing the tasks, assistance would be provided. The moderator also intervened if the participant wandered off task or went too deep into a task or the system crashed. During the experiment, the moderator also encouraged participants to think-aloud during the experiment as it was not natural for the expert participants to talk to themselves. *The think-aloud instrument provided the usability study with the means to gather information about the usability of SERDIF and its potential usefulness.*

The post usability survey. The post usability survey used in this study is the Post-Study System Usability Questionnaire (PSSUQ) to support the think-aloud protocol with a standard quantitative evaluation instrument. The PSSUQ has three versions: version 1 (v1) [Lewis, 1992], version 2 (v2) [Lewis, 2002] and version 3 (v3) [Sauro and Lewis, 2016]. The second version (v2) of the questionnaire was used in this study as it has three more questions (Q3, Q5 and Q13), which referred to the effectiveness and efficiency of the system. Both are considered key aspects to evaluate the usability of SERDIF and to explore the research question of this thesis.

The PSSUQ (v2) is a standard questionnaire meant to assess the usability progress during the development of a system with 19 questions. The standard aspect of the questionnaire allows the comparison of the results with following interactions or similar tools. The PSSUQ follows 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 and 1-19. In this scale, the lower the value, the higher the satisfaction.

The PSSUQ instrument provided the usability study with the means to gather information about the user satisfaction with SERDIF.

5.1.4 Metrics

The metrics gathered to assess the usability and potential usefulness of the framework combined quantitative (summative) and qualitative (formative) metrics to support the findings. The motivation for these metrics was to have at least two metrics for each of the usability characteristics and to balance the limitations of thematic analysis (Section 5.1.5) by using quantitative metrics. The quantitative metrics were also useful to track the usability progress across the three phases using standard post-usability questionnaires. The motivation for the metrics adopted is discussed in detail in (Section 5.1.5).

The quantitative and qualitative metrics are presented below with the name of the metric between parentheses.

Quantitative. The task being completed or abandoned with/out assistance from the usability session moderator (task completion or failure) and the number of assists (assists during task completion); the time spent to complete each task during the usability testing session (time per task); the PSSUQ (usability test survey scores and scales).

Qualitative. The observational findings from the session transcripts with the think-aloud comments of the participants, the open answers the user survey, which allowed users to record text statements for further clarification of the scores; the usability moderator notes that recorded participant thoughts in a summarised manner for an initial impression on the usability, which were not used in the analysis (observational findings).

In this thesis, task completion refers to a successful completion of the task. The time per task metric should be interpreted in combination with the rest of the usability metrics described above to minimise the subjectivity of the results. A participant that spent a long time completing a task may have struggled or enjoyed the task as participants were given sufficient time to explore and complete the tasks. For example, a long time on task associated with many moderator assists and difficulties during the completion of the task can indicate limited usability for a given task and participant.

The usability metrics were associated with the definition of usability including the effectiveness, efficiency and satisfaction attributes, and the potential usefulness of the framework for HDRs, which are summarised in Table 5.1. The **effectiveness was measured** using the assists during task completion, as the moderator's interventions during the Think-Aloud protocol (see Section 5.1.3) can have an effect on the task completion rates [Lewis, 2014]. The **efficiency was measured** using the time spent in completed data linkage tasks. The **satisfaction was measured** using the PSSUQ scores and scales gathered from the participants. The **usability problems and potential usefulness of the framework were identified** by the observational findings from the qualitative data presented above.

Table 5.1: Summary of the usability metrics associated with the usability aspects of the study. The black circle indicates the presence of the element and the blank circle the absence.

Evidence	**Effectiveness**	**Efficiency**	**Satisfaction**	**Usability problems**	**Potential usefulness**
Assists during task completion	●	○	○	○	○
Data linkage time	○	●	○	○	○
Usability test survey	○	○	●	○	○
Observational findings	○	○	○	●	●

5.1.5 Data Analysis

The metrics recorded during the usability testing sessions were analysed following the six steps of a thematic analysis described by Braun and Clarke 2006 [Braun and Clarke, 2006]: step 1: Familiarise yourself with the data; step 2: Generate initial codes; step 3: Search for themes; step 4: Review the themes; step 5: Define and name the themes; step 6: Produce the report.

Thematic analysis is an accessible and structured method to summarise key features shared across qualitative research data while providing a clear manner to report the findings.

The six steps were conducted following the trustworthy thematic analysis approach by Nowell et al. 2017 [Nowell et al., 2017] towards minimising the subjectivity and increasing the credibility of the qualitative analysis. The authors recommended that researchers describe the thematic analysis steps in a precise, consistent and exhaustive manner while making the data and methods available for the readers.

The implementation of the thematic analysis for this thesis is the following.

Step 1: Familiarise yourself with your data. While in a typical thematic analysis only qualitative data (in many formats) is considered, quantitative metrics were also included as they provided further insight in the analysis to support the findings. The quantitative metrics were used in *Step 5: Define and name the themes* and *Step 6: Produce the report* to complement the contextualised results obtained from the qualitative data with generalisable external insights, in order to produce a more complete view of the usefulness of the framework.

The recorded metrics (data) include the previous quantitative and qualitative metrics mentioned in Section 5.1.4, and these are available in the data folder for each phase in the GitHub repository for this thesis to be found at:

https://github.com/navarral/phd-thesis

Data was collected by different interactive means by the usability moderator: stopwatch (time per task), manual typing during the session (moderator notes), automatic process using an online version of the questionnaire (usability test survey scores, scales and open comments), automatic/manual transcription of the audio recordings of the sessions[1] (usability session transcripts), and manual input after the usability session transcription (task completion or failure).

The audios and transcripts were stored locally and then, the audio files were deleted and the transcripts without any personal information uploaded into a Taguette [Rampin and Rampin, 2021] local database for the subsequent coding. The documents are named Participant ID_datetime (e.g. P1_GMT20211118) for easier management and interpretation of the audit trail.

Step 2: Generate initial codes. The interactive data collection influenced the generation of the codes as the moderator and data analyst were the same person (the author of this thesis), but two other researchers, (supervisors of this thesis) were involved in the coding process. Besides the initial influence mentioned above, the researcher did not use an initial codebook at the start of the process, addressing the data with an open mind. The initial generation of the codes was performed line by line and assigning specific codes with a deductive approach. Then, redundant codes were merged, unifying codes assigned to the same type of statements. The number of highlights per code started to show some tendency for some of the codes at this stage. In addition, the code 'open tag' stored sentences which seemed to not fit into any of the other codes but that could be useful in the later iterations of the coding process.

The Taguette software [Rampin and Rampin, 2021] proved to be a great asset for the tagging and categorisation of the sentences in the transcripts, facilitating the researcher's complex task to highlight the sentence, assign a code and write a memo

[1] The transcriptions were transcribed manually for Phase 1 and automatically for Phase 2 and 3

as a description of the code for the debriefing with the other two researchers in *Step 3 Search for themes*. The memo was useful as some of the codes' names would have been hard to be understood by themselves for the other two researchers involved in the analysis and future researchers trying to reproduce this work that are not experts in the content.

At this stage, the three researchers involved in the process met a few times (2-3 times for ∼30min) within a week for a debriefing of the coding process. In each meeting, the researchers tried to reduce the number of codes by merging codes discussing similar concepts. The descriptions associated to the merging codes were rephrased to include all the relevant information from the predeceasing codes. After a minimum of three iterations, the final codebook was agreed by consensus between the three researchers. The codebook generated at each iteration and a document summarising all highlights are available in the GitHub site for the thesis.

Step 3: Search for themes. The searching for themes process was conducted by displaying the number of references per participant for each code as an annotated heat map. The heat map was generated using one of the exports supported in Taguette's software. The export included the transcription highlights per document (i.e. per participant ID) with the associated code as a CSV file (all_tags_date.csv documents in the GitHub of the thesis).

The heat map rows were reordered by number of references but also by their importance in assessing the objective of the usability test for each phase (Section 5.1.1). The reordering naturally led to codes that informed the achievement of the requirements. However, some of the codes did not fit into the requirements, which were left as 'open theme' and not discarded at this step.

Step 4: Review the themes. The document highlights were read again two times to validate the assignment of each code to a particular theme. During the review, the naming and descriptions of the codes had to be adjusted.

In addition, the 'open theme' codes started to shape into different themes regarding new features or functionalities that are not present in the framework, emerging requirements, overall usability experience when conducting the experiment, to name a few.

Step 5: Define and name the themes. The themes identified are described in detail as the findings of the usability study, supported by the quantitative metrics and thematic analysis of the PSSUQ open comments, if enough comments are provided. Therefore, triangulating participant comments, questionnaire responses and observations to demonstrate the findings from multiple perspectives. The metrics are visualised as box plots, stacked bar plots and heatmaps, providing visual context to clarify patterns in the data and reference for each finding. The visualisations showcase the strength of the approach taken in combining quantitative and qualitative data within the definition of the themes. In addition, the data linkage time and PSSUQ scales comparing the previous and current phases of the usability study were also included as input evidence in the generation of findings.

Step 6: Produce the report. The thematic analysis report was written as a usability report since the aim of the analysis was to study the metrics recorded during the usability evaluation. Therefore, the structure of the document was as follows: executive summary, methods, metrics, results, analysis, findings and recommendations, and next steps. The content of the report is included in Sections 5.2, 5.3 and 5.4 together with the recorded metrics in the GitHub of the thesis in order to ensure the transparency of the methods and findings. The study did not finalise by checking the findings with all the participants as it was not viable to organise a meeting where all the expert participants were present and could participate in a discussion. Instead the usability report was presented as part of the agenda in research meetings of the use case projects, where only the experts attending the meeting had the chance to revert to the author with any comments or feedback. Researchers that could not attend the meetings were also given the possibility to revert to the author with any comments or feedback.

The key elements of the thematic analysis for each of the steps are summarised in Table 5.2. The elements include the presence of the qualitative and quantitative metrics, the number of researchers involved in the process and the use of the Taguette software [Rampin and Rampin, 2021] as a tool to facilitate the process execution and documentation.

Table 5.2: Summary of the thematic analysis elements included in each step of the process. The black circle indicates the presence of the element and the blank circle the absence.

Thematic analysis	**Qualitative metrics**	**Quantitative metrics**	**# Researchers**	**Taguette**
Step 1	●	●	I	●
Step 2	●	○	III	●
Step 3	●	○	I	●
Step 4	●	○	III	●
Step 5	●	●	III	●
Step 6	●	●	I	○

5.2 Usability Test - Phase 1

The first usability test of the study is presented in this section. The content of this section has been peer-reviewed, presented and published [Navarro-Gallinad et al., 2021] in the 10th International Joint Conference on Knowledge Graphs (IJCKG 2021, in cooperation with ACM/SIGAI), which is a premium academic forum on Knowledge Graphs. The content has been adapted to follow the structure for dynamic elements presented at the start of this chapter.

5.2.1 Context of Use

Name. Use Case 1 - AAV in Ireland
Description. Anti-neutrophil cytoplasm antibody (ANCA)-associated vasculitis (AAV) is 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 (Fig. 5.2).

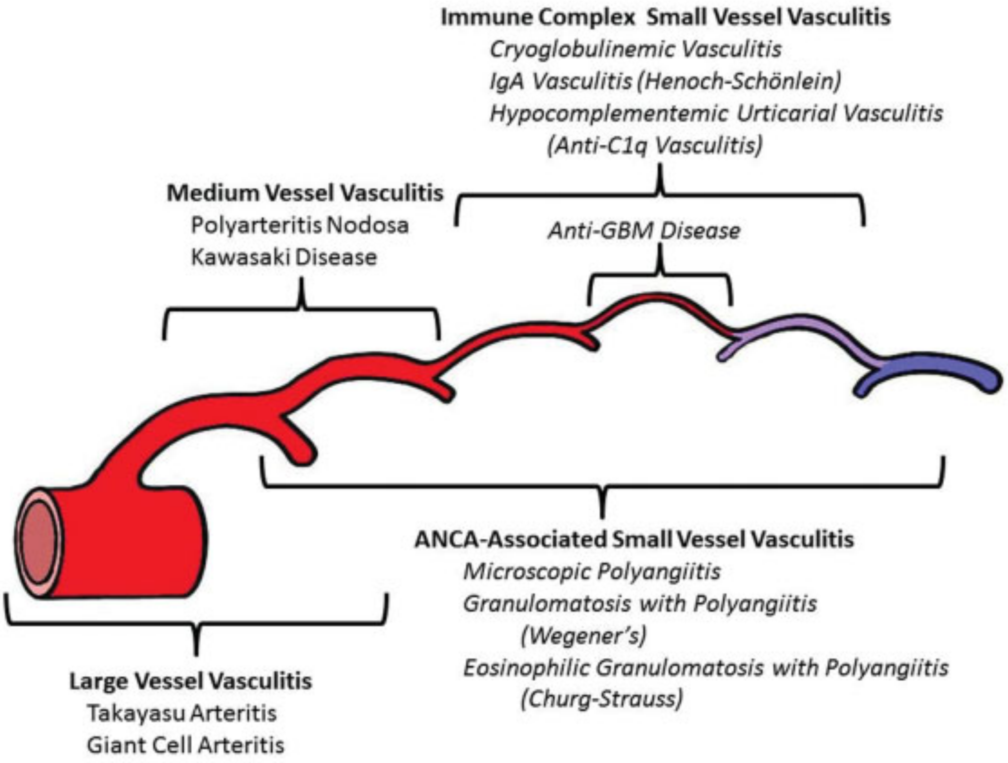

Immune Complex Small Vessel Vasculitis
Cryoglobulinemic Vasculitis
IgA Vasculitis (Henoch-Schönlein)
Hypocomplementemic Urticarial Vasculitis
(Anti-C1q Vasculitis)
Medium Vessel Vasculitis
Polyarteritis Nodosa
Kawasaki Disease
Anti-GBM Disease
ANCA-Associated Small Vessel Vasculitis
Microscopic Polyangiitis
Granulomatosis with Polyangiitis
(Wegener's)
Eosinophilic Granulomatosis with Polyangiitis
(Churg-Strauss)
Large Vessel Vasculitis
Takayasu Arteritis
Giant Cell Arteritis

Figure 5.2: Type of blood vessels inflamed by different forms of vasculitis [Jennette et al., 2013].

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]. The potential environmental triggers include pollutants associated with job exposures, released during natural disasters and through farming, and UV radiation [Scott et al., 2020]. Understanding the environmental trigger could lead to predicting when flares of the disease may occur for individual patients.

Research projects. Two healthcare data linkage projects studying the environmental triggers of AAV were identified for the focus in Phase 1: (i) the HEalth data LInkage for ClinicAL benefit (HELICAL) project HELICAL [2023] and (ii) the AVERT project (AVERT) [AVERT, 2022; Reddy et al., 2019]. HELICAL is a 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. AVERT is an Irish project with the goal of predicting autoimmune risk by investigating the environmental trigger of AAV flares and its prevention. In both projects, standard Knowledge Graphs (KG) technologies are the approach chosen to combine multiple diverse data sources such as patient registries and environmental data by their spatial and temporal common features.

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 standard Knowledge Graphs.

Therefore, the AAV paradigm is an ideal opportunity to apply the SERDIF framework to enable hypothesis validation of the environmental triggers for this disease. The 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.

5.2.2 Expert User Requirements

An initial set of requirements were gathered from expert researcher meetings and through undertaking a consensus process with HDRs in the preliminary evaluation (Section 3.3.2).

The expert user requirements evaluated in Phase 1 are the following:

Requirement 1 (R1.0). *Enable HDRs to query specific clinical patient data to retrieve linked environmental data, without the need for knowledge of the underpinning semantic web technologies.*

Requirement 2 (R2.0). *Support the understanding of the use and limitations of the*

linked environmental data to support identification of flares for rare diseases.
Requirement 3 (R3.0). *Export selected clinical and environmental data to be used as input in statistical models for data analysis.*

5.2.3 Initial Framework Implementation

The initial implementation of SERDIF is presented in this section for **Use Case 1 - AAV in Ireland**. The initial framework was developed as a result of a state-of-the-art review presented in this thesis (Section 3.1 and 3.2 and the positive outcome of the usability evaluation conducted on an initial UI (Section 3.3). The initial framework was a combination of three components: a methodology, a knowledge graph and a user interface.

5.2.3.1 Methodology Component

The methodology started as a series of steps that should be taken by the researcher to define and use the necessary spatiotemporal data structures to combine clinical and environmental data. The methodology was divided into six main processes illustrated in Fig. 5.3.

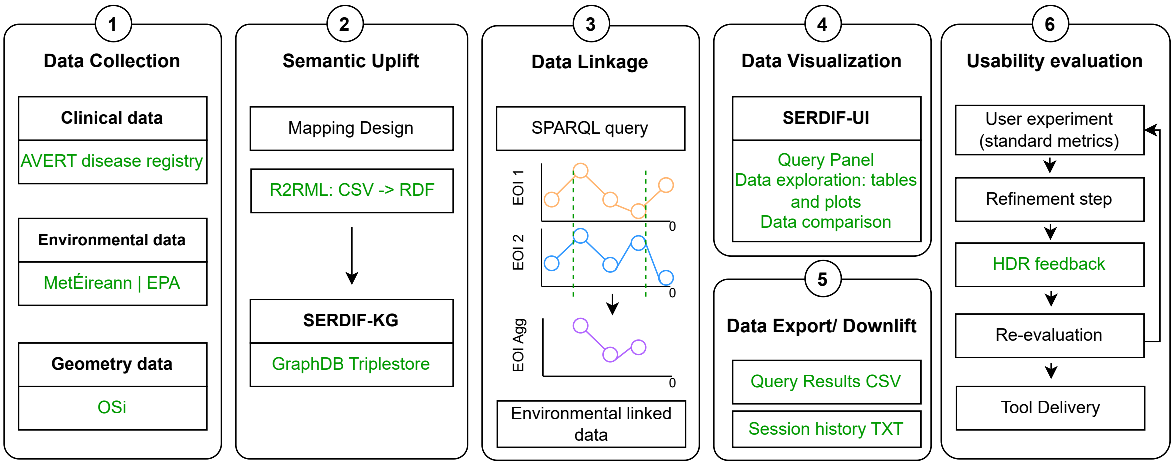


Figure 5.3: Diagram of the SERDIF methodology with the in-use application in green evaluated in Phase 1.

Step 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.

Step 2: Semantic uplift. This process designs a declarative mapping to uplift the environmental data gathered from the data collection process to RDF. 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 [Perry et al., 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.

Step 3: Data linkage. This process defines a spatiotemporal query as a SPARQL template. A SPARQL query template that has placeholders (or variables) for users' input (*Step 4: Data visualisation*) 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.

Step 4: Data visualisation. 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-centred, focused on domain experts' requirements, to develop an effective tool. The initial requirements can be extracted from expert consensus within a project.

Step 5: Data export/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 one 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.

Step 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. UI) will be ready to be delivered.

5.2.3.2 Knowledge Graph Component

The KG component benefits from the spatiotemporal data structures to combine clinical and environmental observations through locations, from geometry data; and relative periods from the clinical events. The KG was developed as a result of implementing the first two steps of the SERDIF methodology, *Step 1: Data collection*, *Step 2: Semantic uplift* and *Step 3: Data linkage.*

Step 1: Data collection (implementation). Clinical, environmental and geometry data were manually collected (or accessed in the case of clinical data) in the data collection process. Previous work from the AVERT project [AVERT, 2022] 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 [OSI, 2023]. The interest from HDRs was the validation of environmental triggers for AAV; therefore, environmental data was gathered from land-based stations within the country. In the first iteration, weather [MET, 2023] and pollution [EPA, 2023] data are collected as tabular files (CSV). In addition, metadata files that include the environmental variables descriptions and station locations for each data source were also gathered.

Step 2: Semantic Uplift (implementation). This process designed a declarative mapping to uplift the environmental data gathered from the data collection process to RDF. The semantic uplift process provided 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-used the Sensor Network (SOSA) existing vocabulary [Haller et al., 2017] facilitating spatiotemporal reasoning due to the organisation levels (Listing 5.1): geolocated samplers (*sosa:Sample*r) that include samplings (*sosa:Samplings*) as time series data. Observation values were 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 was 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 was based on the initial requirements gathered from expert consensus [Navarro-Gallinad et al., 2020].

Listing 5.1: An example of a SERDIF sampler data structure diagram.

```
# -- Namespaces --------------------
@prefix sosa: http://www.w3.org/ns/sosa/
@prefix serdif: http://serdif.org/kg/datasource/
@prefix serdif-epa-station: http://serdif.org/kg/datasource/pollution/
   EpaAirQDataHly/
@prefix geo: http://www.opengis.net/ont/geosparql#
@prefix xsd: http://www.w3.org/2001/XMLSchema#
# -- Sampler example ---------------
serdif-epa-station:EPA-75 a sosa:Sampler, geo:Feature ;
   serdif:stationCode "EPA-75" ;
   geo:hasGeometry serdif-epa-station:EPA-75_-7.6996_52.3547 ;
   sosa:madeSampling serdif-epa-station:EPA-75_-7.6996_52.3547
      _2013-12-02T000000Z ;
[...]
.
# -- Sampler geometry --------------
serdif-epa-station:EPA-75_-7.6996_52.3547 a geo:Geometry ;
  geo:asWKT  "POINT(-7.6996 52.3547)"^^geo:wktLiteral .
# -- Sampling ----------------------
serdif-epa-station:EPA-75_-7.6996_52.3547_2013-12-02T000000Z a sosa:Sampling
    ;
  sosa:resultTime "2013-12-02T00:00:00Z"^^xsd:dateTime ;
  serdif:hasO3 "41.0"^^xsd:float ;
  serdif:hasTemperature "5.1"^^xsd:float ;
  [...]
.
[...]
```

Regarding the implementation, R2RML-F was 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 was 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 were imported to a GraphDB triplestore [Ontotext, 2022]. The R2RML mappings and ontologies used are made available in the GitHub for the thesis to reproduce the uplift process:

https://github.com/navarral/phd-thesis/ (implementation/)

Table 5.3: Data sources summary for Step 2 of the methodology implementation in P1.

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	28k

Table 5.3 summarises 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 included a validation step that checks for any syntax errors which stops on error. The triplestore was chosen due to the GeoSPARQL support, key for HDR queries, and their easy to use interface to develop applications.

Step 3: Data linkage (implementation). In this process, the clinical and environmental datasets could have been 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 (Listing 5.2). The SPARQL queries used in Phase 1 are made available in the GitHub for the thesis to reproduce data linkage process:

https://github.com/navarral/phd-thesis/ (implementation/)

Location. The RDF clinical graph included 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 was necessary due to the missing explicit triple pattern shared between both data sources (i.e. the point geometries do not concur). GeoSPARQL functions [Perry et al., 2012] enabled spatial reasoning between geometries with functions like *geof:distance* or *geof:sfWithin*. In this case, *geof:sfWithin* was the function chosen since the aim was 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 contained events such as disease activity and remission state dates or hospital admissions represented as *xsd:dateTime* data types (?dateEvent). Hence, environmental observations were filtered for a specific period related to the clinical events. In this case, the period was defined by the lag from the event (?dateLag) and the duration (?dateStart), which in Listing 5.2 is of 7 and 30 days respectively.

Listing 5.2: Spatiotemporal reasoning used in the SERDIF querying process as a SPARQL example.

```
# -- Namespaces ---------------------------------------
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)
```

5.2.3.3 User Interface Component

The User Interface (UI) component was designed from a user-centred perspective to support HDRs access, explore and export the linked health-environmental data with appropriate visualisations, and by facilitating the query formulation for expert users. The UI was developed as a result of implementing *Step 4: Data visualisation* and *Step 5: Data export/downlift* steps of the SERDIF methodology.

Step 4: Data visualisation (implementation). An initial UI was designed with features to understand the environmental linked data such as summaries, data tables and plots, for HDRs. Dash [Plotly, 2023] is the Python framework used to build the UI, which contains query input and main panels in a coordinated view. For a better understanding of the data visualisation step an example UI was made available at:

https://w3id.org/serdif/

Query input panel. The user can select multiple options as input from the clinical data to retrieve the associated environmental data in this panel (Fig. 5.4A). The input options were dynamically displayed using live predefined queries executed while the UI was running, providing a flexible visual tool. If new data became available as the research progressed, the UI would be able to adapt to the new data automatically. The query inputs were 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 enabled the submit button at the bottom of this panel when all the options were selected and data is available. When the user clicked on the submit button, the selected options were 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 UI consisted of three tabs: home, comparative and query number (Fig. 5.4B). The home tab provided an introduction to the UI use together with acknowledgement of the data sources combined throughout their website links. In addition, an interactive choropleth map was available to explore the number of samplers per county. After each query submission, a new tab was 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 displayed the raw data outcome resulting from the query as a heat map data table. The plot sub tabs (see Fig. 5.4C and D) were interactive and allowed the user to visually explore the data table normalised 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 facilitated comprehension of the complex relationship between environmental variables and wind.

Previous queries could be visually compared in the comparative tab by specific variable (see Fig. 5.4E). In addition, the queries could be arranged into groups to potentially reveal signals that the individual queries were hiding.

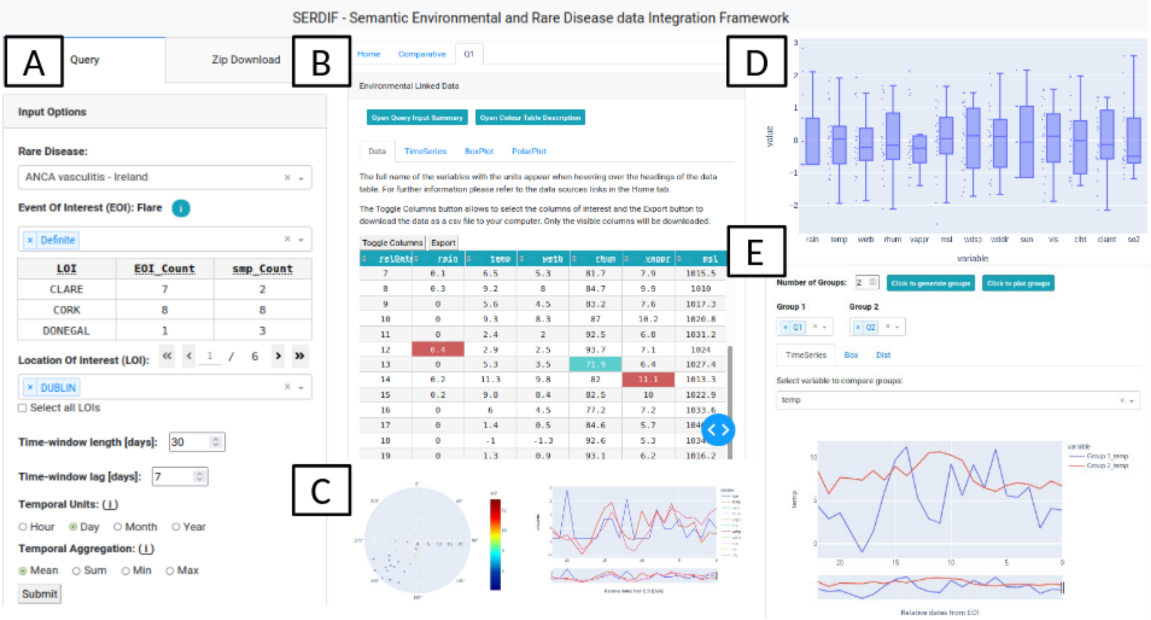
Step 5: Data export/downlift (implementation). The query number tab generated after each query submission contains an export button situated on top of the data table (Fig. 5.4B). The user could click the export button to download the selected columns from the data table as a CSV file. Moreover, the user could 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 was located in the ZIP Download tab from the Query input panel (Fig. 5.4A).

5.2.4 Evaluation Against the Requirements

This section includes the usability test sample size, participant tasks, evaluation results, analysis and conclusions for Phase 1.

5.2.4.1 Sample Size

The initial sample size estimation is based on a target value of likelihood of discovering a problem and the chances of this problem to occur, computed using the cumulative



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

Location Of Interest (LOI): « ‹ 1 / 6 › »

× 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	184
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 5.4: Screenshot of the SERDIF UI displaying (A) the query input panel, which allows domain 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 visualisations, (C) polar, time series and (D) box plots; and (E) the comparison tab, where previous queries can be compared in groups.

binomial probability formula [Lewis, 2014]. The target for this usability study was to discover >90% of the usability problems that can happen 25% of the time. The estimated sample size for this is n=10 [Lewis, 2014], which will be the starting point for the progressively increasing pool of expert participants.

A pool of 10 HDRs currently researching the environmental triggers of AAV in Ireland were recruited for Phase 1. The HDRs were international professors, researchers and PhD students with fluent English, who are analysing AAV clinical data for Ireland in their research. The expert participants were analysed only for being a researcher in the field rather than their experience or expertise within the field.

5.2.4.2 Participants Tasks

The expert participants were asked to complete the following tasks, designed to assess the three user requirements (Section 5.2.2). The tasks were derived from consensus among HDRs with real workflows in mind and are presented in Table 5.4.

Table 5.4: Tasks that the participants are asked to complete during the usability testing associated with the requirements from Section 5.2.2 in Phase 1.

Task	**Requirement**	**Description**
T1	Querying	Read the text within the 'Home' tab in the main panel aloud and explore a data source of your choosing by right clicking and opening it in a new tab. Once you feel you understand where the data comes from please click the button 'Open map' to visualise the density of data points within the Locations of Interest (LOI). You will be done when you mention a few of the LOIs that have data points available.
T2	Querying	Submit a query using the input options available in the 'Query' tab on the left most panel. While you are selecting the query options, please explain aloud the specific choices that you make. As you make choices new ones will become available for you to select. Once you reach the end, please click the 'Submit' button once, which will only be available if valid inputs have been selected. You will be done when a new tab named 'Q1' is displayed in the main panel of the UI. Please only submit one query.
T3	Understanding	Explore the 'Q1' tab. In order to understand the table underneath there are two buttons that help you interpret the data table that you see. click both buttons on top, 'Open Query Input Summary' and 'Open Colour Table Description', commenting aloud on their usefulness and clarity as explanations. Then, read the text in the 'Data' sub tab and talk through the interpretation of the variables and cell background colours (red and blue) of the data table. You can select the columns that you want to display with the 'Toggle Columns' button. Note that you can scroll vertically and horizontally and use the page arrows (on the right bottom corner) to check the amount of data that you queried by scrolling. You will be done when you feel you have sufficiently explored this table.

T4	Understanding	Explore the other tabs beside the data tab. All three plots (time series, box plot and polar plot) are interactive with features for different purposes, please use those features for each plot while mentioning the usefulness of each one in understanding the environmental data. You will be done when you have explored through the interactive features for each of the three plot tabs.
T5	Understanding	Create two new queries. After you have submitted them following the same instructions in Task 2. You will see 'Q2', 'Q3' as new tabs on the main panel. Now you can compare them in the 'Comparative' tab. Read the text in this tab and select the number of groups you wish to make. Then click on 'Click to generate groups' and new inputs per group will appear underneath. Arrange your queries in the groups and click on 'Click to plot groups'. Now, new sub tabs will have appeared with three different visualisations. Please mention aloud the usefulness of each one in a similar manner to Task 3.
T6	Exporting	Choose one of the queries to export ('Q1', 'Q2' or 'Q3') and go to that tab. Click the 'Export' button from the data table to download the data as a csv file. Finally, open the 'Zip Download' tab on the left side panel and click on 'Download Zip with all datatables' to download all data for all the queries submitted during the session, as a zip file. You will be done when understanding where this data is stored.
T7	Summary	Finally, could you summarise verbally your overall experience in completing these tasks using the SERDIF UI? Then, please proceed to complete the PSSUQ questionnaire.

5.2.4.3 Results

The usability evaluation results for Phase 1 were gathered during the implementation of *Step 6: Usability evaluation* step of the SERDIF methodology.

Quantitative.

The quantitative results of Phase 1 include task completion (Fig. 5.5), time per task (Fig. 5.6) and the PSSUQ scores and scales (Fig. 5.7).

Task completion. Most of the tasks were completed successfully in Phase 1 (60 out of 63, Fig. 5.5). 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 visualisation of the comparative plots. Even though the sample size for this experiment was 10, the task completion evidence (i.e. session transcript) was only available for 9 participants due to a technical error during the recording of P4. The participants required assistance from the moderator in the tasks associated with the data linkage process (T1-6) with a mean of 25 assists per participant (**Effectiveness benchmark**). Less assistance was required in T7 where the participants had to summarise the overall experience in using SERDIF.

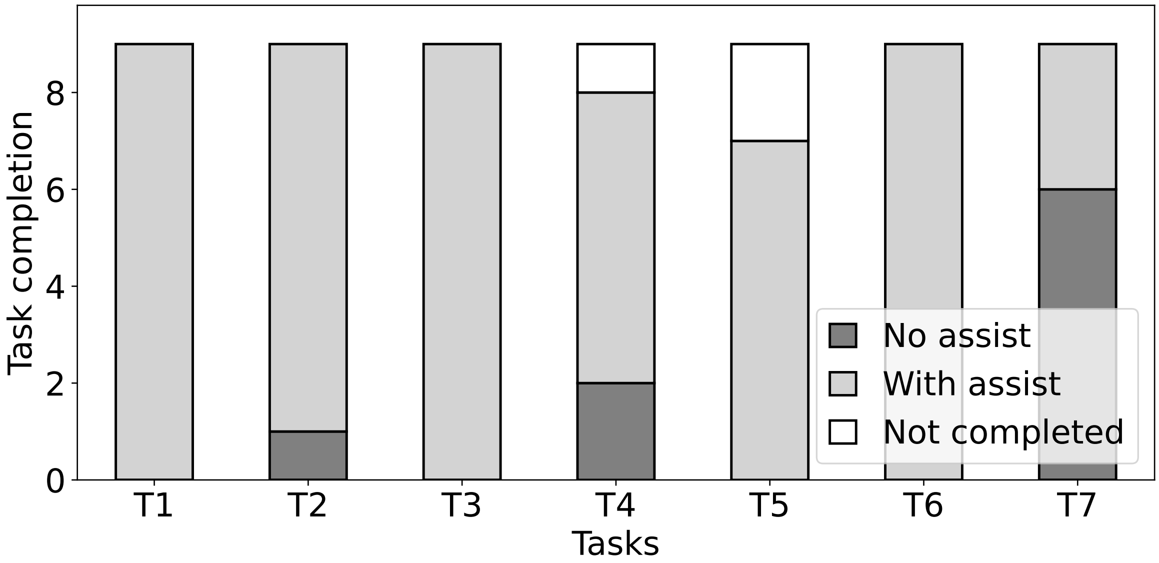


Figure 5.5: Task completion bar plot for Phase 1 including the if assistance was needed from the participants (n=9): no assistance required (no assist, dark grey), with at least one intervention to assist (with assist, light grey) and task or subtasks not completed (not completed, white).

Time per task. The box plots in Fig. 5.6 compared participants' time spent on tasks in Phase 1. The tasks followed an increasing trend from T1-T5, regarding the querying and understanding of the linked data, and dropped for the last two tasks T6, where the resulting linked data was exported, and T7, where the overall experience using SERDIF was summarised. More specifically, the tasks T1 and T6, regarding querying (*Requirement 1 - Query*) and exporting the linked data (*Requirement 3 - Export*) respectively, had an InterQuartile Range (IQR) (i.e. length of the boxes) under 3min. Tasks T4-5, representing (*Requirement 2 - Understand*), had an IQR over 3min. The participants spent more time understanding the linked data in a more heterogeneous manner than formulating the query or exporting the linked data. Overall, the data linkage tasks (T1-6) had a mean of 40min (**Efficiency benchmark**).

PSSUQ scores and scales. Most of the box plots in Fig. 5.7 for the PSSUQ scores had a median of 2 (15 out of 19 questions) and an IQR below 2 points (16 out of 19

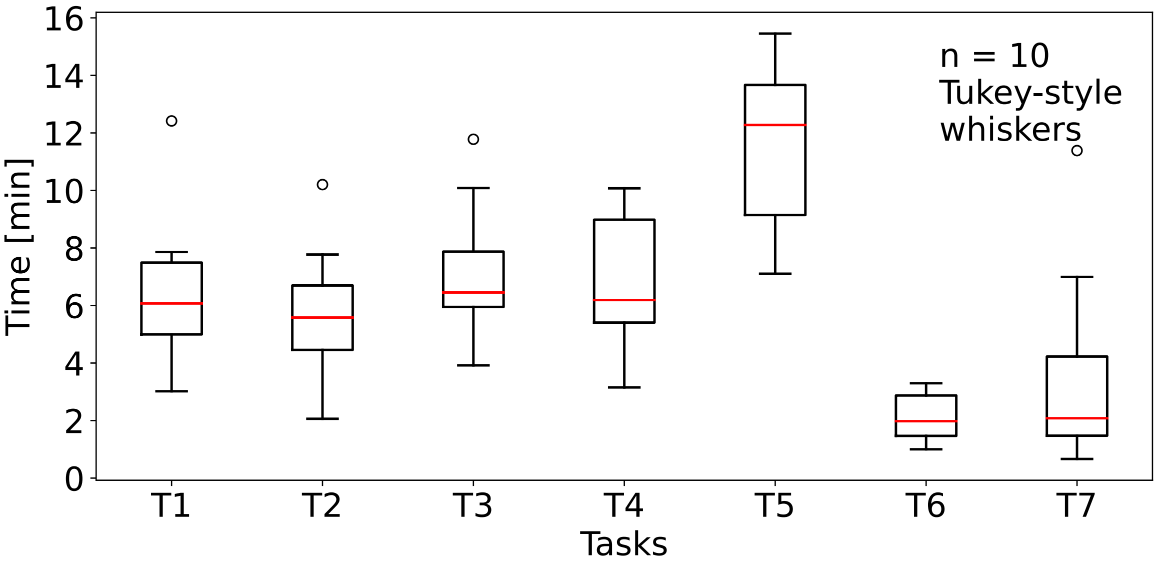


Figure 5.6: Time spent magnitude and variability to complete each task during the usability session represented as a box plot for Phase 1.

questions). Error messages (Q9), expected capabilities and functions (Q18) and overall satisfaction (Q19) were the worst scores per question; becoming productive (Q8) and information organisation (Q15) represented the best scores. The PSSUQ scales with the average scores SysUse, InfoQual, IntQual and Overall had similar values with SysUse being slightly better than the rest and IntQual being the most dispersive (i.e. larger IQR). The Overall scale had a median of 2.21 (**Satisfaction benchmark**), where 1 is the best possible score and 7 the worst.

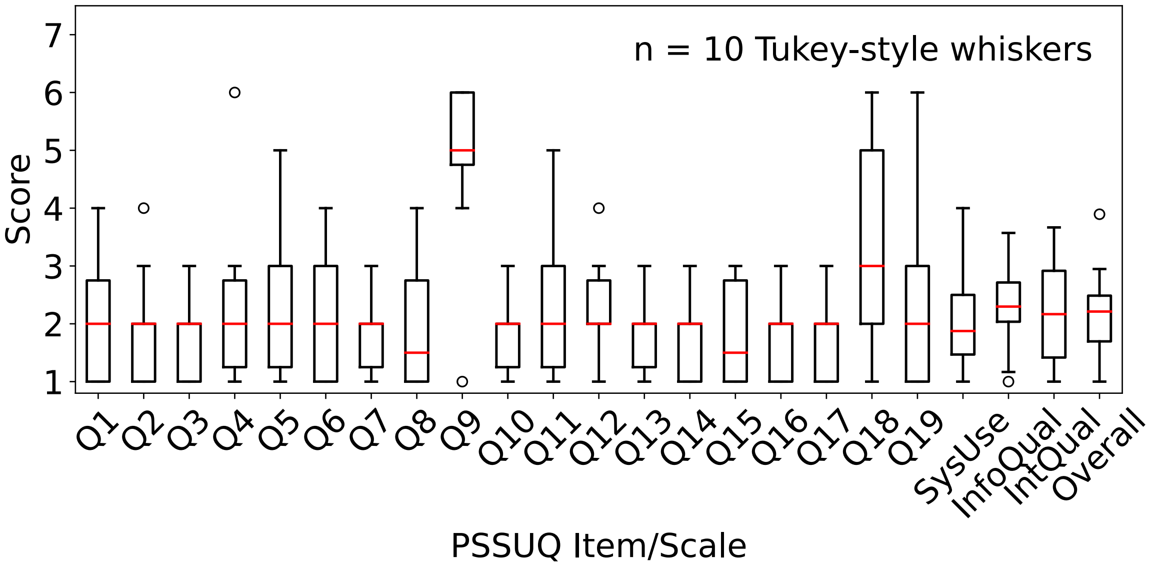


Figure 5.7: PSSUQ scores box plots for Phase 1 with the four averaged metrics (SysUse, InfoQual, IntQual and Overall) on the right end with sample sizes of 80, 70, 30 and 190. The scores are in a Likert 7 points scale where the lower the value the higher the satisfaction.

Qualitative

The qualitative results included the transcriptions of the think aloud comments of the participants, the notes taken by the usability moderator during the experimental sessions and the PSSUQ open comments. The transcripts and notes are available in the GitHub repository of this thesis:

https://github.com/navarral/phd-thesis/ (evaluation/phase-1/)

Few expert participants (3 out of 10) provided comments in the PSSUQ open comments space. The comments are presented below but they had not been included for consideration in any case in the thematic analysis as evidence for the themes.

Participant 7 (P7). *"I didn't get any error messages, so I'm not sure how to answer this..."*

Participant 8 (P8). *"The visual plots could be improved. It would be nice to look at each environment variable separately. For example: it would be nice to have a time series or map plot showing raw values for temperature or pollution across space and time. Regardless, the platform provides all that you need to access environmental data and is easy to use."*

Participant 9 (P9). *"I didn't have enough knowledge on proposed variables so the benefits of using this application increases as there is more information about the variables." "As there are some written instructions, I could easily follow how to use the system." "I didn't receive any error message, however, the buttons such as "submit" did not get activated until all the information entered." "If there was some information about the goals of each task, it would be easier to understand."*

5.2.4.4 Thematic Analysis

The text transcriptions from the usability sessions were analysed following the six step process of thematic analysis previously outlined in Section 5.1.5. The results of the analysis are presented as a summary of the themes (Table 5.5), a stacked bar plot with the types of assists required during the usability sessions (Fig. 5.8) and a heat map that presents the codes leading to the themes for traceability of the results (Fig. 5.9). The table and figures are presented as a reference for the evidence statements for the findings of this usability test in Phase 1.

Themes and findings summary. The themes name, findings, number of times a code has been referenced within the theme (references) and scope of the finding in terms of the SERDIF components and experimental methodology are summarised in Table 5.5.

The themes captured the usability and potential usefulness of SERDIF when facilitating the process of linking environmental and health data, the need to refine the starting requirements with a more specific and transparent data linkage process, while also including a mention of the complexity of some of the features of the UI, as well as the testing methodology issues encountered when completing the tasks using the UI.

Table 5.5: Thematic analysis summary of the usability sessions transcripts colour coded as in Fig. 5.9 heat map.

Themes	Findings	References	Scope
Useful approach for linking data	Positive overall user experience emphasising the data exploration features and the usefulness of SERDIF	126	Framework
Requirements refinement	The origin and processing of the linked data is unclear and the environmental data needs to be linked for a period prior to the flare events.	69	Framework
Complex text and features	Some of the plots are complex and the technical jargon makes text descriptions hard to understand.	75	UI
Testing methodology	The task's wording, delays and control malfunctioning during the virtual experiment session reduced the overall usability of SERDIF.	269	Testing

Types of moderator's assistance. The total number of moderator assists was first gathered from the transcripts of the usability sessions. Then, the assists were sorted into three categories based on the component that required the assistance: the system is crashed or is not responsive (System issue), the task is confusing or not completed in a sequential manner (Task complex) and the design and/or content of the user interface complicates the completion of the tasks (Navigation and content complex). Furthermore, the assists were once more divided into the task they related to, providing more context for the following analysis (Fig. 5.5).

Codes of the thematic analysis. The resulting codes and themes are presented as an annotated heat map. The heat map allows for a more detailed and transparent view of the codes associated with each of the participants. For example, the influence of each participant (i.e. the code distribution) on the total number of references can be checked for further insights when proving the evidence for each code. The ordering of the rows was based on the number of references within a theme (Section 5.1.5). The descriptions of the codes for this phase are made available in Appendix D.

Findings. The author of this thesis observational findings from the Phase 1 usability test (thematic analysis supported by the quantitative metrics) are described in this section together with the recommendations generated for the next phase.

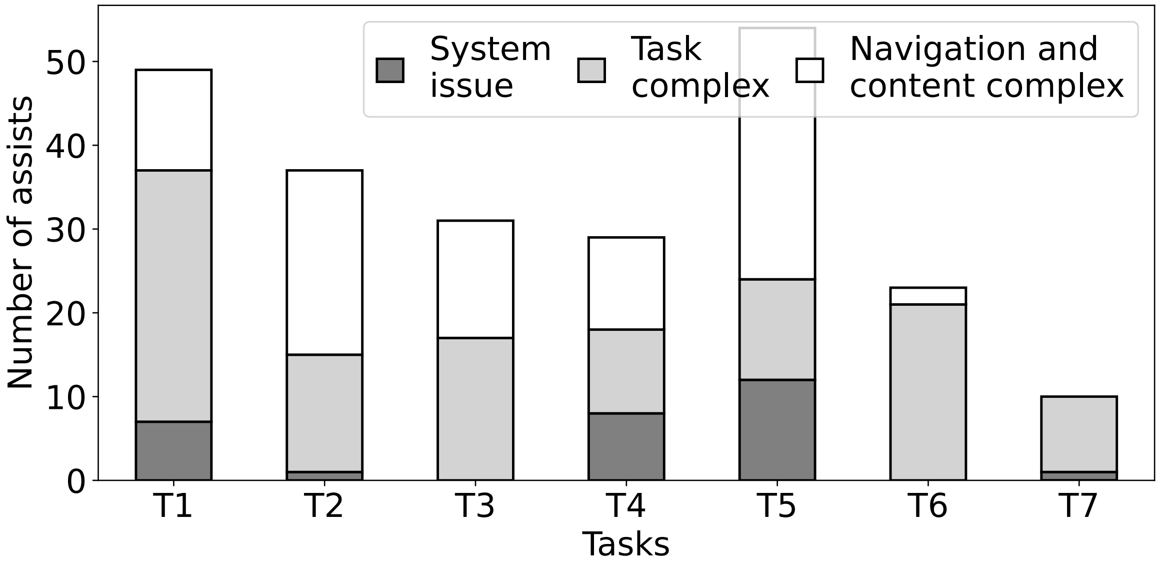
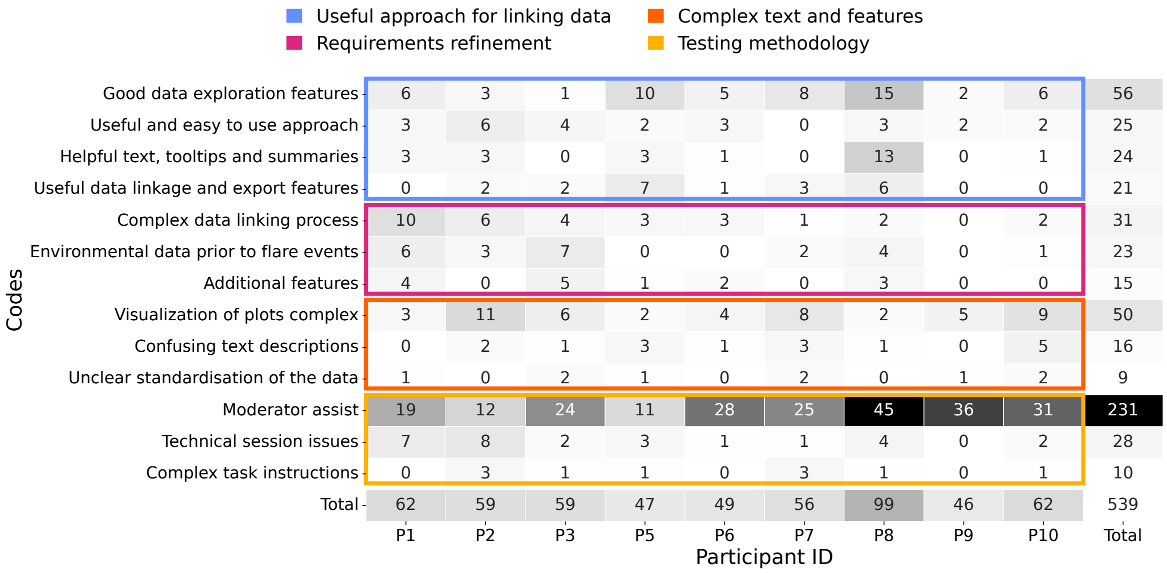


Figure 5.8: Stacked bar plot with number of assists from the usability testing moderator to the participants (n=9) in Phase 1. The type of assistance is included as: the system is crashed or is not responsive (System issue, dark grey), the task is confusing or not completed in a sequential manner (Task complex, light grey) and the design and/or content of the user interface complicates the completion of the tasks (Navigation and content complex, white).

Theme: Useful approach for linking data

Finding (P1.F1). *"Positive overall user experience emphasising the data exploration features and the usefulness of SERDIF."*

Evidence. The expert participants were able to complete the tasks with assistance from the moderator (60 out of 63, Fig. 5.6), which indicates that the environmental and health data could be linked using SERDIF. The approach taken with SERDIF was useful for researchers conducting health-environmental studies as it facilitated the complex task of linking data using a query panel with clear options, good data exploration features with helpful descriptions and plots, and useful data export features (126 out of 539, Table 5.5 and Fig. 5.9; and Q18 score and SysUse scale in Fig. 5.7). The second most frequent code overall, supported by the best PSSUQ score in Q15, was towards the data exploration approach taken with the UI, including summaries, data tables with coloured cells for high and low values and multiple plots to visualise complementary dimensions of the data (56 out of 539, Fig. 5.9; and Q15, Fig. 5.7). Moreover, the PSSUQ scales were lower than the norm defined for the PSSUQ version 2 [Lewis, 2002], the lower the value the higher the satisfaction; and provided a reference for the next versions of SERDIF.



Codes	P1	P2	P3	P5	P6	P7	P8	P9	P10	Total
Good data exploration features	6	3	1	10	5	8	15	2	6	56
Useful and easy to use approach	3	6	4	2	3	0	3	2	2	25
Helpful text, tooltips and summaries	3	3	0	3	1	0	13	0	1	24
Useful data linkage and export features	0	2	2	7	1	3	6	0	0	21
Complex data linking process	10	6	4	3	3	1	2	0	2	31
Environmental data prior to flare events	6	3	7	0	0	2	4	0	1	23
Additional features	4	0	5	1	2	0	3	0	0	15
Visualization of plots complex	3	11	6	2	4	8	2	5	9	50
Confusing text descriptions	0	2	1	3	1	3	1	0	5	16
Unclear standardisation of the data	1	0	2	1	0	2	0	1	2	9
Moderator assist	19	12	24	11	28	25	45	36	31	231
Technical session issues	7	8	2	3	1	1	4	0	2	28
Complex task instructions	0	3	1	1	0	3	1	0	1	10
Total	62	59	59	47	49	56	99	46	62	539

Figure 5.9: Codes and themes that emerged from the thematic analysis of the usability sessions transcripts in Phase 1 as a heat map for traceability.

Recommendations:

- P1.F1.1. Keep with the current data linkage approach with a UI that makes the linkage process intelligible and the data outcome explorable with data visualisations.

Theme: Requirements refinement

Finding (P1.F2). *"The origin and processing of the linked data is unclear and the environmental data needs to be linked for a period prior to the flare events."*

Evidence. While the participants were able to complete the tasks presented during the experiment, they required assistance because of the complexity of the navigation and content almost half of the time (91 out of 231, Fig. 5.8). Furthermore, most of these assists were associated with the understanding the use and limitations of the linked data (T2-5, *Requirement 2 - Understanding*) after the linkage query (53 out of 91, Fig. 5.8). The limitation of SERDIF towards *Requirement 2 - Understanding* was also supported by the time taken to complete tasks T2-5 (Fig. 5.7) and the worse PSSUQ scores for the questions related to having all the expected capabilities and overall satisfaction (Q18 and Q19, Fig. 5.7). The rest of the assists were related to tasks T1 (12 and 2 out of 91, Fig. 5.9) and T6 (2 out of 91, Fig. 5.8) implying that *Requirement 1 - Querying* needed more refinement than *Requirement 3 - Exporting.*

This affirmation was also supported by the time per task for tasks T1 and T7 (Fig. 5.6).

The most referenced codes for this second theme indicated that the requirements had to be refined towards being more specific (*Requirement 1 - Querying*) and transparent (*Requirement 2 - Understanding*) about the data linkage process (54 out of 69, Fig. 5.9). The other code in this theme referred to the additional features that would improve SERDIF, such as extending the query options (*Requirement 1 - Querying*) with more aggregation methods, health event categories and custom inputs (15 out of 69, Fig. 5.9).

Recommendations:

- P1.F2.1. Specify in *Requirement 1 - Querying* that environmental data needs to be associated with individual health events through location and time, within the region of the event and a period of data before the event.

- P1.F2.2. Include in *Requirement 2 - Understanding* provenance and lineage metadata to support the understanding of the linked data.

- P1.F2.3. Include in *Requirement 2 - Understanding* data protection risk information to provide evidence of the compliance with GDPR now that the linked data refers to individuals (see P1.F2.1)

- P1.F2.4. Update SERDIF components to meet the refined requirements.

Theme: Complex text and features

Finding (P1.F3). *"Some of the plots are complex and the technical jargon makes text descriptions hard to understand."*

Evidence. The complexity of linking environmental and health data was clear from the expert participants' total comments (75 out of 539, Fig. 5.9). Furthermore, the majority of the code references to this theme identified the types and elements of plots complex (50 out of 75, Fig. 5.9). The plots were presented in Fig. 5.4C, D, E to provide an example of what a participant encountered while completing tasks T4-6. The time per task increase (Fig. 5.6) and the number of assists (Fig. 5.8) required peaked when the participants were asked to use the plots to compare two queries in T5, the most complex task for the participants. The highest number of assists in T5 was attributed to the navigation and content being complex (30 out of 54, Fig. 5.8).

The rest of the codes in this theme referred to the technical jargon used in the text (16 out of 75, Fig. 5.9) and the standardisation of the raw values in the data table to highlight low and high values (9 out of 75, Fig. 5.9).

Recommendations:

- P1.F3.1. Remove the comparative tab.
- P1.F3.2. Update the plots to represent environmental data linked to individual health events instead of aggregated data.
- P1.F3.3. Clarify the complex jargon in the text descriptions.
- P1.F3.4. Present a visible and concise description of the data standardisation process instead of a button to unfold the information.

Theme: Testing methodology

Finding (P1.F4). *"The task's wording, delays and control malfunctioning during the virtual experiment session reduced the overall usability of SERDIF."*

Evidence. The task wording generated some confusion for the expert participants making the tasks complex to complete. This resulted in the expert participants requesting assistance due to the complexity of the task's wording (113 out of 231, Fig. 5.8). Overall, all the participants needed some type of assistance as denoted by the most referenced code (231 out of 539, Fig. 5.9). The rest of the codes in this theme referred to the technical issues encountered while completing the tasks (28 out of 269, Fig. 5.9) and to the explicit comment on the complexity of the tasks by the participants (10 out of 269, Fig. 5.9). The technical issues effect on the testing methodology were also supported by the time per task IQR over 3min (Fig. 5.6) and the system issues present in the number of assists in tasks T1, T4 and T5 (Fig. 5.8). Testing methodology was also the most referenced theme in Phase 1 (269 out of 539, Table 5.5 and 5.9).

Recommendations:

- P1.F4.1. Rephrase the tasks using easy-to-understand language.
- P1.F4.2. Divide complex tasks into smaller subtasks.
- P1.F4.3. Make the UI accessible through an online hosting service.

– P1.F4.4. Build a graph database with example health data (see P1.F2.1).

5.2.4.5 Conclusions

The first iteration of SERDIF that was evaluated with HDR participants involved in researching **Use Case 1 - AAV in Ireland** yielded an encouraging outcome (P1.F1). First, the methodology to integrate environmental data with longitudinal and geospatial diverse clinical data was suggested to be potentially useful for HDRs for this case study. Second, the associated developed knowledge graph structure was effective in linking graph data through a SPARQL query. Third, the initial SERDIF UI allowed researchers to access, explore and export the linked health-environmental data.

The summative usability part of the evaluation set the benchmarks for the usability metrics measuring the effectiveness, efficiency and satisfaction for the following phases of the study. The formative usability part was successful in deriving findings and recommendations for the further development of the SERDIF framework. The requirements needed to be refined towards clarifying the data linkage process of individual health events with environmental data (P1.F2). The methodology needed to include new elements to enhance the transparency of the linkage process while complying with the General Data Protection Regulation law (GDPR) [EU, 2016] for processing individual environmental-patient linked records in the next phase. The KG and UI components of the framework needed update as a consequence of the changes in the methodology. Some of the UI features and text description required clarification as they were complex to understand by the participants (P1.F3). Furthermore, an alternative method to conduct the usability experiment needed exploration to address the delay and control malfunctioning of the current remote control functionality of the video conferencing platform (P1.F4).

5.3 Usability Test - Phase 2

The second usability test of the study is presented in this section. The content of this section has been peer-reviewed and published [Navarro-Gallinad et al., 2022] in the Semantic Web Journal (in cooperation with IOS Press Content Library), which is a top venue for topics on Semantic Web technologies. The content has been adapted to follow the format presented at the start of this chapter.

5.3.1 Context of Use

Name. Use Case 2 - Kawasaki Disease (KD) in Japan
Description. Kawasaki Disease (KD) is a rare vasculitis of unknown aetiology and the

second main cause of acquired heart disease in children around the world. While KD is around the world, Japan is the country with the highest incidence (Fig. 5.10). The current theory sustains that an unidentified agent enters through the upper respiratory tract and causes a dramatic immunologic response, in certain genetically predisposed children younger than 5 years old [Rife and Gedalia, 2020; Rowley and Shulman, 2018]. The pathogenesis theory for KD is supported by the apparent seasonality of KD reported in countries in Asia, North America and Europe [Uehara and Belay, 2012]. Furthermore, climatological studies point towards an environmental agent transported by tropospheric winds to be the trigger link of this paediatric vasculitis [Rodó et al., 2016,1].

Research projects. The WINDBIOME project [ISGlobal, 2023] was added to the research projects from **Use Case 1 - AAV in Ireland**. The WINDBIOME project aims to discover the etiological agent of KD towards the development of an early warning system for healthcare institutions and citizens. Researchers in this project are trying to link epidemiological national survey data from KD in Japan with the physical, chemical, and biological characteristics of air masses. The air data has been collected at different spatial and temporal scales and with different equipment. Researchers using the heterogeneous datasets would benefit from an effective data linkage method where the origin and processing of the data could be easily tracked with provenance metadata. That is why researchers from WINDBIOME are open to apply emerging KG approaches to manage and link the data in graph databases as an alternative to accessing the datasets in a shared and secured repository.

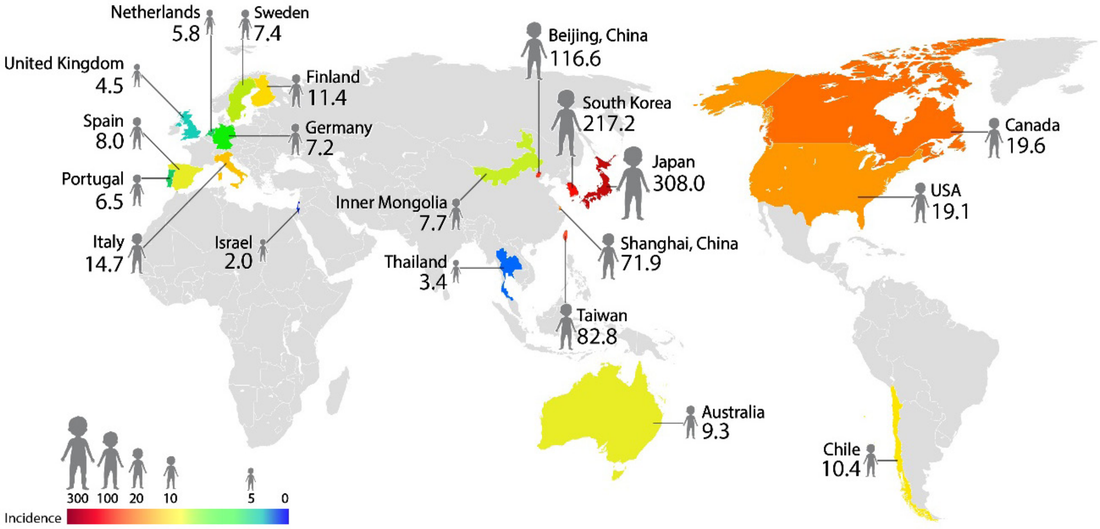


Figure 5.10: Graphical summary of the incidence of Kawasaki disease (KD) from [Kim, 2019].

For that reason, KD in Japan was an ideal use case to apply the SERDIF framework, in supporting HDRs in hypothesis validation of environmental agents. The use case also tested whether the framework would be flexible and useful for HDRs with a different research culture, and where population health data is linked to multiple environmental data at a regional level rather than a country level.

5.3.2 Expert User Requirements

The expert user requirements were refined after the analysis of the evaluation results in Phase 1 (Section 5.2.4). The refinement included (R1.1) a more specific definition of the type of health data to be linked with environmental data on how to define the link so it is relevant to individual health events (P1.F2.1); (R2.1) the addition of metadata to trace the origin and processing of the linked data towards a reuse of the data (P1.F2.2) that complies with data protection regulations (P1.F2.3); (R3.1) the extension of export options for publication of machine-understandable data as RDF graphs towards mandatory Open Science practices including FAIR data publication (EU project requirement) [Commission, 2022].

The expert user requirements evaluated in Phase 2 are the following with the underlined words denoting the refinements of the requirements:

Requirement 1 (R1.1). *Enable HDRs to query environmental data associated with individual health events through location and time, within the region of the event and a period of data before the event.*
Requirement 2 (R2.1). *Support the understanding of event-environmental linked data and metadata, with its use, limitations and data protection risk for individuals.*
Requirement 3 (R3.1). *Export event-environmental linked (meta)data to be used as input in statistical models for data analysis (CSV) and for publication (RDF).*

5.3.3 Framework Implementation

The SERDIF components have been updated after the results from Phase 1 of the usability study (P1.F1.1 and P1.F2.4, Section 5.2.4) and the refined requirements (Section 5.3.2). The updates are presented here for each of the components of the framework (KG, Methodology and UI) and for the usability testing execution. Providing enough context to interpret the evaluation results of this section.

5.3.3.1 Methodology Component

The methodology is a series of steps that guides the researcher in linking particular events with environmental data using SW technologies. The *Step 4: Data visualisation* and *Step 5: Data export/Downlift* of the SERDIF methodology have been expanded to include information regarding the data protection and privacy, and for open data publication, respectively.

Step 4: Data visualisation (update). This process now includes relevant information on the origin of the data and the processing steps performed to link the health and environmental data (P1.F2.2). Data protection and privacy information was also included in this step to help the user in making a more informed decision on how the personal data (e.g. health data) is processed based on the requirements of the contract signed to use the data under a specific purpose (e.g. a data sharing agreement or consent form) and complies with the General Data Protection Regulation (GDPR) [EU, 2016] (P1.F2.3).

Step 5: Export/downlift (update). This process has been extended to include exporting the linked data as RDF. The user is now provided with the semi-automatic means to make the data interoperable, for later publication in an open data repository as Findable, Accessible, Interoperable and Reusable (FAIR) [Wilkinson et al., 2016]. The publication of the data will only be possible after explicit permission from the data controller, and when it does not provide means for re-identification of the data subject (P1.F2.3). The data protection information is represented in RDF using the Data Protection Vocabulary (DPV) [Pandit, 2022; Pandit et al., 2019].

5.3.3.2 Knowledge Graph Component

The KG component is where environmental and health data is linked together through location and time using RDF and SPARQL queries. The implementation of *Step 1: Data collection*, *Step 2: Semantic uplift* and *Step 3: Data Linkage* was updated to address the researchers' requirements of Phase 2 (Section 5.3.2).

Step 1: Data collection (implementation). The air pollution data source was replaced with the European Environmental Agency (EEA) [EEA, 2022] while the weather data remains the same from MetEireann. The uplifted data includes 25 weather and 2000 air quality individual datasets with hourly data from the 2000-2021 period and for the Republic of Ireland.

The clinical data from AVERT project [AVERT, 2022] stored in a separated triplestore was replaced with simulated health data (Listing 5.3) (P1.F4.4). This implementation update granted the possibility to include researchers that had not signed a data sharing agreement to use and process the health data in the following steps. The simulated health data was made available in a different repository within the triplestore (i.e. internal repository) to work as a proof of concept for a federated scenario, where health data does not leave the storage location but it is only consulted.

Listing 5.3: Snippet of a health event uplifted to RDF Turtle format.

```
PREFIX serdif: <https://serdif.adaptcentre.ie/kg/2022>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
# -- Health event data ------------------
serdif:event-42 a prov:Activity ;
# -- Type of event ----------------------
    rdfs:label "Definite"@en ;
    rdfs:comment "Definite flare event for patient 4"@en ;
# -- Individual with the event ---------
    prov:wasAssociatedWith  serdif:ID-4 ;
# -- Time ------------------------------
    prov:startedAtTime "2013-12-01T00:00:00Z"^^xsd:dateTime ;
# -- Location --------------------------
    prov:atLocation serdif:ID-4-geo .
# -- Geometry --------------------------
serdif:ID-4-geo a prov:Location, geo:Feature ;
    geo:hasGeometry [
        geo:asWKT "POINT(-6.3132 53.1131)"^^geo:wktLiteral ] .
# -- Individual with the event ----------
serdif:ID-4 a prov:Agent ;
    rdfs:label "ID-4"@en ;
    rdfs:comment "Individual with ID-4"@en .
```

Step 2: Semantic uplift (implementation). Only the minimum health event data to enable the linkage was uplifted to RDF, in particular the name, description, time and location of the health event using the provenance ontology (PROV-O) [Lebo et al., 2013] (Listing 5.3).

Environmental data was described using the RDF Data Cube vocabulary (QB) [Cyganiak and Reynolds, 2014] instead of the Semantic Sensor Network Ontology (SOSA/SSN) [Haller et al., 2017] that was used in the KG at the start of P1. QB focuses on describing statistical and multi-dimensional datasets, which presents an advantage as it provides a general data structure and flexibility to represent other types of environmental data besides sensors such as occupational exposures or environments. The datasets can be represented as a time series of observations with a fixed location (GeoSPARQL geometry) using *qb:Slice*. The slice structure facilitates the access to subsets of data and allows for metadata to be included at the slice level. The environmental data described using QB has been published as open data in Zenodo

[Navarro-Gallinad, 2021]. The total amount of triples uplifted for environmental data has increased from 30M to 80M triples in Phase 2. The R2RML mappings and ontologies used in Phase 2 are made available in the GitHub for the thesis to reproduce the uplift process:

https://github.com/navarral/phd-thesis/ (implementation/)

Step 3: Data linkage (implementation). The data is linked using a SPARQL query [29] reasoning over location and time. The GeoSPARQL function *geof:sfWithin* [Perry et al., 2012] is used to select the environmental datasets within the region of the event (e.g. a county or country). The *xsd:dateTime* [Biron and Malhotra, 2004] data type allows for the selection of a certain period before the health events (Listing 5.2). Environmental data is associated with health data for a particular region and period before the health event, as the researchers are trying to understand the risk factors that led to the event (P1.F2.1). The SPARQL queries used in this step were updated to construct health-environmental linked data as in Listing 5.4. The query linked environmental data for each of the events as a *qb:Slice*, allowing the information to be retrieved at the individual event level. The researchers had the option to aggregate the events if necessary at the data analysis step of their workflows. The SPARQL queries used in Phase 2 are made available in the GitHub for the thesis to reproduce data linkage process:

https://github.com/navarral/phd-thesis/ (implementation/)

Listing 5.4: SPARQL query elements to construct a health-environmental dataset as a RDF graph.

```
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX serdif: <https://serdif.adaptcentre.ie/kg/2022>
PREFIX sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
# -- Construct an environmental-health dataset ------
CONSTRUCT{
   ?sliceName a qb:Slice;
      qb:sliceStructure     serdif-slice:sliceByTime ;
      serdif-dimension:event   ?eventRef ;
      serdif-dimension:area ?eventGeo
      qb:observation     ?obsName .
   ?obsName a qb:Observation ;
      qb:dataSet      ?datasetName ;
```

```
        sdmx-dimension:timePeriod  ?obsTimePeriod ;
        ?envProp      ?envVarV .
}
# -- Aggregate environmental data per event ---------
SELECT ?event ?yearT ?monthT ?envProp ?lag (AVG(?envVar) AS ?envVarV)
# -- Unite individual results to construct slices ---
{ serdif:event-42 } UNION { ... }
```

The KG for Phase 2 is publicly available by making the GraphDB triplestore available as a Linked Data (LD) SPARQL endpoint:

https://w3id.org/serdif/

5.3.3.3 User Interface Component

The UI component is designed from a user-centred perspective to support HDRs access, explore and export the linked health-environmental data with appropriate visualisations, and by facilitating the query formulation for domain experts. The implementation of *Step 4: Data visualisation* and *Step 5: Data exporting/downlift* was updated to address the researchers' requirements of Phase 2 (Section 5.3.2).

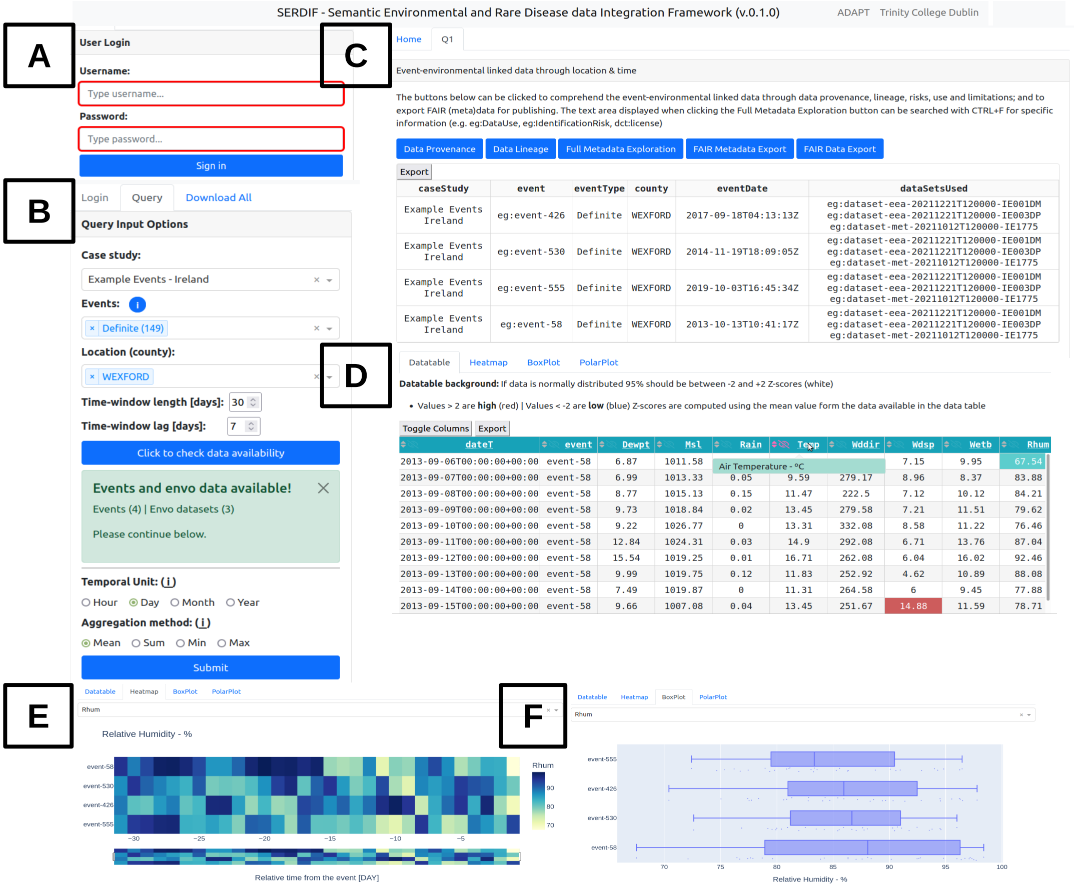
Step 4: Data visualisation (implementation). The UI was updated to present environmental data linked to individual health events instead of aggregated data for an event type (P1.F3.2). For a better understanding of the data visualisation step an example UI is made available at:

https://w3id.org/serdif/

The updates are associated with existing features of the UI like the data table (Fig. 5.11D) and the time series plot, which became a temporal heat map (Fig. 5.11E) (P1.F3.2). The comparative tab was removed as it was confusing to the users and the relevance was not clear at the end of Phase 1 (P1.F3.1). The text descriptions in the main panel and the query generated tabs were clarified and simplified to facilitate the understanding of the linkage process, content (P1.F3.3) and standardisation of the environmental values in the data table (Fig. 5.11C, D) (P1.F3.4).

In addition, a check button was added for the user to check if environmental data is available for options selected from the previous dropdowns (Fig. 5.11B); and a login step was added to avoid multiple participants running queries at the same time, which could affect the experimental metrics (Fig. 5.11A).

The updated KG granted SPARQL queries with the option to gather metadata information beyond the weather and air pollution variables descriptions with the dataset



SERDIF - Semantic Environmental and Rare Disease data Integration Framework (v.0.1.0)

User Login

Username: Type username...

Password: Type password...

Query Input Options

Case study: Example Events - Ireland

Events: Definite (149)

Location (county): WEXFORD

Time-window length [days]: 30

Time-window lag [days]: 7

Events and envo data available!

Events (4) | Envo datasets (3)

Please continue below.

Temporal Unit: (i)

○ Hour ● Day ○ Month ○ Year

Aggregation method: (i)

● Mean ○ Sum ○ Min ○ Max

Event-environmental linked data through location & time

The buttons below can be clicked to comprehend the event-environmental linked data through data provenance, lineage, risks, use and limitations; and to export FAIR (meta)data for publishing. The text area displayed when clicking the Full Metadata Exploration button can be searched with CTRL+F for specific information (e.g. eg:DataUse, eg:IdentificationRisk, dct:license)

caseStudy	event	eventType	county	eventDate	dataSetsUsed
Example Events Ireland	eg:event-426	Definite	WEXFORD	2017-09-18T04:13:13Z	eg:dataset-eea-20211221T120000-IE001DM eg:dataset-eea-20211221T120000-IE003DP eg:dataset-met-20211012T120000-IE1775
Example Events Ireland	eg:event-530	Definite	WEXFORD	2014-11-19T18:09:05Z	eg:dataset-eea-20211221T120000-IE001DM eg:dataset-eea-20211221T120000-IE003DP eg:dataset-met-20211012T120000-IE1775
Example Events Ireland	eg:event-555	Definite	WEXFORD	2019-10-03T16:45:34Z	eg:dataset-eea-20211221T120000-IE001DM eg:dataset-eea-20211221T120000-IE003DP eg:dataset-met-20211012T120000-IE1775
Example Events Ireland	eg:event-58	Definite	WEXFORD	2013-10-13T10:41:17Z	eg:dataset-eea-20211221T120000-IE001DM eg:dataset-eea-20211221T120000-IE003DP eg:dataset-met-20211012T120000-IE1775

Datatable background: If data is normally distributed 95% should be between -2 and +2 Z-scores (white)

- Values > 2 are **high** (red) | Values < -2 are **low** (blue) Z-scores are computed using the mean value form the data available in the data table

dateT	event	Dewpt	Msl	Rain	Temp	Wddir	Wdsp	Wetb	Rhum
2013-09-06T00:00:00+00:00	event-58	6.87	1011.58				7.15	9.95	67.54
2013-09-07T00:00:00+00:00	event-58	6.99	1013.33	0.05	9.59	279.17	8.96	8.37	83.88
2013-09-08T00:00:00+00:00	event-58	8.77	1015.13	0.15	11.47	222.5	7.12	10.12	84.21
2013-09-09T00:00:00+00:00	event-58	9.73	1018.84	0.02	13.45	279.58	7.21	11.51	79.62
2013-09-10T00:00:00+00:00	event-58	9.22	1026.77	0	13.31	332.08	8.58	11.22	76.46
2013-09-11T00:00:00+00:00	event-58	12.84	1024.31	0.03	14.9	292.08	6.71	13.76	87.04
2013-09-12T00:00:00+00:00	event-58	15.54	1019.25	0.01	16.71	262.08	6.04	16.02	92.46
2013-09-13T00:00:00+00:00	event-58	9.99	1019.75	0.12	11.83	252.92	4.62	10.89	88.08
2013-09-14T00:00:00+00:00	event-58	7.49	1019.87	0	11.31	264.58	6	9.45	77.88
2013-09-15T00:00:00+00:00	event-58	9.66	1007.08	0.04	13.45	251.67	14.88	11.59	78.71

Figure 5.11: Screenshot of the SERDIF UI displaying (A) the login panel, (B) the query input panel with selected options, (C) the tabs generated after submitting a query, which includes metadata section with the FAIR export button, (D) a data table and three different visualisations, (E) heat map, (F) box plot.

descriptors Albertoni et al. [2020], and the origin and processing of the datasets [Lebo et al., 2013] (Listing 5.4). Information about the origin of the data and processing steps was made available through the data provenance and lineage buttons as RDF graphs (Fig. 5.11C). The user can now explore the full metadata generated for the linkage process (P1.F2.2 and P1.F2.2) that includes dataset descriptors (e.g. licence, distribution, temporal and spatial information and structure of the dataset), data provenance and lineage information (e.g. datasets used and SPARQL query to link the data), data protection and privacy aspects of the linked data (e.g. data controller, processing risks and purpose of the linkage) and the data use (e.g. for studying a specific disease by researchers with a signed data sharing agreement to process the health data).

Step 5: Data export/downlift (implementation). The results from a query can be exported as interoperable data with the potential for researchers to make it FAIR and to further link the data with other studies and for open data publication (Listing 5.5 and 5.6, see [Navarro-Gallinad et al., 2023] for a complete export example).

Listing 5.5: Snippet of a Turtle RDF file describing the metadata of the linkage process between health related events and environmental data.

```
# -- Data Set -------------------------------------------
serdif:dataset-ee-20211008T120000 a qb:DataSet, geo:Feature, prov:Entity,
    dcat:Dataset ;
    dct:title "Air pollution and climate data associated with multiple events
       "@en ;
    dct:description "The dataset is an example result of associating air
       [...]"@en ;
    dct:identifier "https://doi.org/10.5281/zenodo.5544257"^^xsd:anyURI ;
    dct:hasVersion "20211008T120000" ;
    dct:issued "2021-10-08T12:00:00Z"^^xsd:dateTime ;
    dct:publisher <https://www.adaptcentre.ie/>, <https://www.tcd.ie/> ;
    dct:license <https://creativecommons.org/licenses/by-sa/4.0/> ;
  # -- Themes describing the dataset --------------------------------
    dcat:theme <https://www.wikidata.org/entity/Q932068> , [...] ;
  # -- External data sets used to construct this data set ------------
    dct:hasPart <http://example.org/ns#dataset-eea-20211012T120000-IE003AP>,
       [...] ;
  # -- Spatial descriptors -------------------------------------------
    dct:Location geohiveCounty:2ae19629-1454-13a3-e055-000000000001 ;
  # -- Temporal descriptors ------------------------------------------
    dcat:temporalResolution  "P1D"^^xsd:duration ;
    dct:temporal eg:dataset-ee-20211012T120000-temporal ;
  # -- RDF Data cube structure ---------------------------------------
    qb:structure eg:dataset-ee-20211012T120000-dsd ;
  # -- Activity that constructed the data set ------------------------
    prov:wasGeneratedBy eg:agg-dataset-ee-20211012T120000 ;
  # -- Data protection aspects ---------------------------------------
    dpv:hasDataController <https://www.tcd.ie/> ;
  [...]
.
# -- Agents -------------------------------------------------------------
<https://orcid.org/0000-0002-2336-753X> a prov:Person, prov:Agent, dpv:
    DataProcessor .
<https://www.adaptcentre.ie/> a dct:Agent .
<https://www.tcd.ie/> a dct:Agent .
# -- Data provenance and lineage ----------------------------------------
serdif:agg-dataset-ee-20211012T120000-QT-2021-11-24T16%3A16%3A20.590Z
  # -- Type of activity ----------------------------------------------
  a prov:Activity, prvt:DataCreation ;
  # -- External data sets used in the activity -----------------------
  prov:used     <http://example.org/ns#dataset-eea-20211012T120000-IE003AP>,
       [...] ;
  # -- ORCID for the agent that performed the activity ---------------
  prov:wasAssociatedWith  <https://orcid.org/0000-0002-2336-753X> ;
  # -- Activity explanation for humans -------------------------------
  rdfs:comment    "The activity describes a SPARQL query to associate [...]"
     @en ;
  # -- Query to construct the data set -------------------------------
  prvt:usedGuideline [  a prvt:CreationGuideline, prvt:SPARQLquery, sp:
     Construct ;
     sp:text """ CONSTRUCT { ... } WHERE { ... } """ ;
  ] ;
.
```

Listing 5.6: Snippet of a Turtle RDF file with health related events linked with environmental data.

```
# -- Namespaces ---------------------
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX serdif: <https://serdif.adaptcentre.ie/kg/2022>
PREFIX sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
# -- Observations -------------------
# Event 42
serdif:dataset-ee-20211012T120000-IE-event-42-slice a qb:Slice ;
    qb:sliceStructure serdif:sliceByTime;
    serdif:refArea geohiveCounty:2ae19629-1454-13a3-e055-000000000001 ;
    serdif:refEvent serdif:event-42 ;
    qb:observation  serdif:
       dataset-ee-20211012T120000-IE-event-42-obs-20131203T010000Z, [...] ;
.
serdif:dataset-ee-20211012T120000-IE-event-42-obs-20131203T010000Z
    a qb:Observation ;
    qb:dataSet serdif:dataset-ee-20211012T120000-IE ;
    sdmx-dimension:timePeriod "2013-12-02T00:00:00Z"^^xsd:dateTime ;
    serdif:hasO3 "41.0"^^xsd:float ;
    serdif:hasTemp "5.1"^^xsd:float ;
    serdif:hasVappr "9.7"^^xsd:float ;
    [...]
.
```

5.3.4 Evaluation Against the Requirements

This section includes the usability test sample size, participant tasks, evaluation results, analysis and conclusions for Phase 2.

5.3.4.1 Sample Size

The pool of 10 HDRs from **Use Case 1 - AAV in Ireland** was increased with 7 more from this use case, giving a total of 17 HDRs for Phase 2. The additional 7 researchers of this use case are international professors, researchers, PhD students and lab technicians with fluent English, who are analysing KD epidemiological data for Japan in their research.

The increased pool of 17 expert participants enhances the chance to discover >90% of the usability problems that can happen 15% of the time, compared to the 25% occurrence in Phase 1 [Lewis, 2014].

5.3.4.2 Experimental Setup Update

The participants experienced some delays and control malfunctions during the P1 of the usability study. That was because the moderator had to open the UI in a local computer, share the screen during the video conference call and then give remote control to the participant. In Phase 2, a UI was made available directly for the participants under the following URL (P1.F4.3):

https://w3id.org/serdif/

5.3.4.3 Participants Tasks Update

As a result of the findings from Phase 1, the tasks were rephrased using simpler language (P1.F4.1) and were divided into subtasks (P1.F4.2) to improve the readability and to avoid having to go back and forth between the PDF document with the tasks and the browser window with the UI (Table 5.6).

Table 5.6: Tasks that the participants are asked to complete during the usability testing associated with the requirements from Section 5.3.2 in Phase 2.

Task	Requirement	Description
T1	Querying	Read the 'Home' tab information. (a) Was the information presented enough for you to get an overall understanding of the SERDIF framework? (b) If not, what information would you add?
T2	Querying	Login with the credentials provided in the invitation email.
T3	Querying	Submit a query using the input options available in the 'Query' tab on the left panel. (a) Please explain and justify aloud the specific choices that you make. (b) Are the query input options clear to you? (c) Would you add any extra information to get a better understanding? (d) After completing all the inputs, please click the 'Submit' button once.
T4	Understanding	Explore the 'Q1' tab generated after submitting a query. (a) Read the information provided in the introductory paragraph. (b) Explore and discuss aloud the usefulness of the information displayed for each of the three buttons below, in order to understand the event-environmental linked data. (c) Click the following buttons: Data Provenance, Data Lineage and Full Metadata Exploration. (d) Can you say aloud how many data sets were used for each event in the Data Provenance table? (e) Can you identify aloud the data use comment (*eg:DataUse*), identification risk comment (*eg:IdentificationRisk*) and license (*dct:license* [DCMI, 2020]) in the Full Metadata Exploration text?

T5	Understanding	Explore the data table displayed underneath the metadata buttons. (a) Can you understand all the column headings when hovering over the abbreviations? (b) Do you understand the meaning of the cell background colours? (c) Can you hide columns by using the eye icon next to the column heading or by using the 'Toggle Columns' button? (d) After coming to conclusion on the previous questions, discuss aloud whether you found the data table useful to comprehend the event-environmental linked data? (e) Say aloud if you would add any feature to the current display
T6	Understanding	Explore the Heat map, Box Plot and Polar Plot tabs. (a) Select an environmental variable from the inputs provided. (b) Please explain aloud the usefulness and clarity of each plot. (c) Say aloud if you would add any more visualisations to the current ones.

T7	Understanding	Submit one more query with different input options from the ones used before. (a) Which individual event uses the most data sets between Q1 and Q2? (b) Do both data sets (Q1 and Q2) have the same license? (c) Which of the data tables (Q1 and Q2) seem to have more extreme values (i.e., more colored cell backgrounds)? (d) Choose one of the following plots from Heat map, Box Plot and Polar Plot tabs and compare aloud Q1 and Q2 plots. (e) Say aloud if you would add any other feature to help you comprehend better the data at this point.
T8	Exporting	Export/download the metadata and data from the first query (Q1) (a) FAIR data and metadata. (b) Data provenance and Data table as CSV files. (c) Export all data tables at once in the 'Download All' tab in the left panel. (d) Say aloud whether such event-environmental linked data and metadata would be useful and re-usable as input for environmental research.
T9	All	Could you summarise verbally your overall experience in completing these tasks using the SERDIF UI?

5.3.4.4 Results

The usability evaluation results for Phase 2 were gathered during the implementation of *Step 6: Usability evaluation* step of the SERDIF methodology updated for Phase 2.

Quantitative

The quantitative results of Phase 2 include task completion (Fig. 5.12), time per task (Fig. 5.13), the PSSUQ scores and scales (Fig. 5.14) and the usability progress between Phase 1 and 2 comparing the data linkage time and PSSUQ scales (Fig. 5.15).

Task completion. Most of the tasks were completed successfully by the 17 expert participants (146 out of 153, Fig. 5.12). Five participants could not complete T7 due to their query selections in T3 and a coding error which affected the visualisation of the comparative plots. At least one assist was required for almost half of the participants in tasks T3 (*Requirement 1 - Query*), T5-7 (*Requirement 2 - Understand*) and T8 (*Requirement 3 - Export*). The task that required the most assistance was T4, where participants had to explore the provenance metadata to understand the origin and processing of the linked data. The participants did not need assistance when asked to summarise the overall experience in using SERDIF.

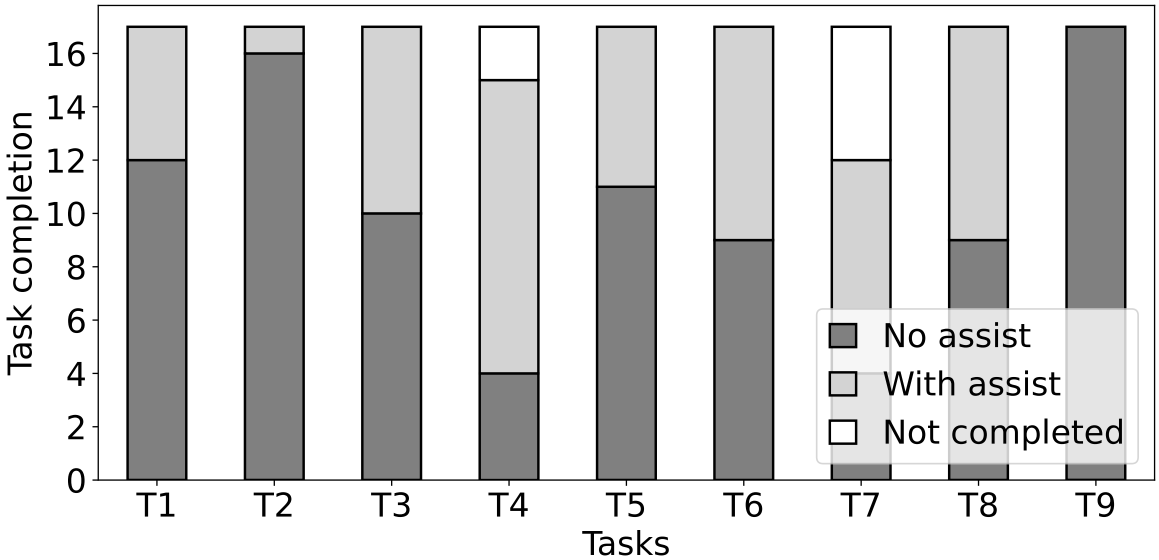


Figure 5.12: Task completion bar plot for Phase 2 including the if assistance was needed from the participants (n=17): no assistance required (no assist, dark grey), with at least one intervention to assist (with assist, light grey) and task or subtasks not completed (not completed, white).

Time per task. The box plots in Fig. 5.13 compared participants' time spent on tasks in Phase 2. The tasks follow an increasing trend from T1-T7, regarding the querying and understanding of the linked data, and drop for the last two tasks T8, where the resulting linked data is exported, and T9, where the overall experience using SERDIF is summarised. The positive trend has a maximum during the course at T4, with a median of 10 min and an IQR of 10 min, being the task that took the longest but also the most dispersive. Most of the tasks (6 out of 9) had an IQR between 2-3min. The participants spent more time understanding the linked data in a more heterogeneous manner (*Requirement 2 - Understand*) than formulating the query (*Requirement 1 - Query*) or exporting the linked data (*Requirement 3 - Export*) as in Phase 1. Overall, the data linkage tasks (T1-8) had a mean of 41min.

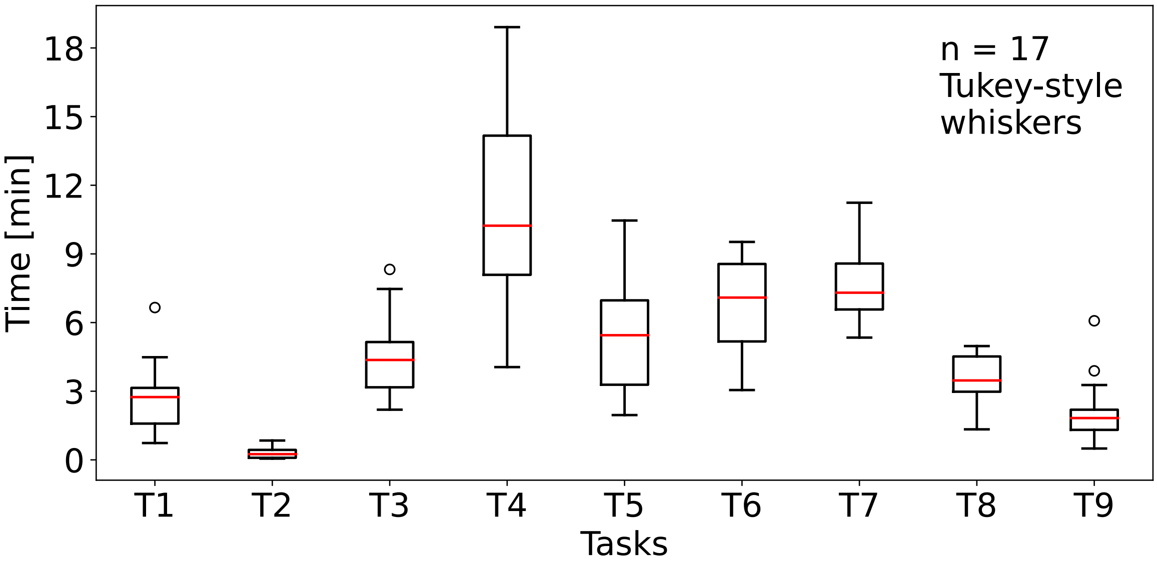


Figure 5.13: Time spent magnitude and variability to complete each task during the usability session represented as a box plot for Phase 2.

PSSUQ scores and scales. Most of the box plots in Fig. 5.14 for the PSSUQ scores had a median of 2 points (13 out of 19 questions) and an IQR of 1 point (16 out of 19 questions). Error messages (Q9) had the worst score with a median of 4 points; and becoming productive (Q8), recovering easily from mistakes (Q10), information being effective (Q14), pleasant (Q16) and likeable interface (Q17) had the best scores with a median of 1 point. The scores with more dispersion were the ones related to completing the tasks quickly (Q4) and efficiently (Q5), and finding the information easily (Q12). The PSSUQ scales with the average scores SysUse, InfoQual, IntQual and Overall had similar values with IntQual being slightly better and less dispersive than the rest. The Overall scale had a median of 1.89, where 1 is the best possible score and 7 the worst.

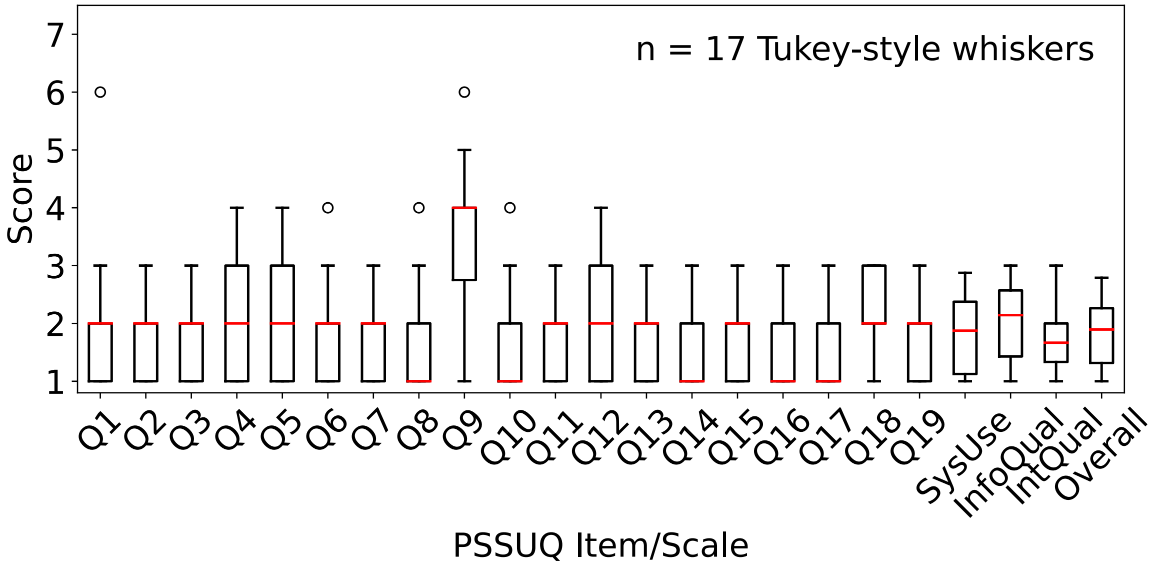


Figure 5.14: PSSUQ scores box plots for Phase 2with the four averaged metrics (SysUse, InfoQual, IntQual and Overall) on the right end with sample sizes of 136, 119, 51 and 323. The scores are in a Likert 7 points scale where the lower the value the higher the satisfaction.

Usability progress between Phase 1 and 2. The progress of usability metrics (efficiency, effectiveness and satisfaction) is presented in Fig. 5.15. The efficiency remained similar between both phases with median around 40 min for the data linkage tasks (Fig. 5.15A). The effectiveness improved in Phase 2 as denoted by the lower number of assists required per participants when completing the data linkage tasks (Fig. 5.15B). The satisfaction identified from the PSSUQ scales has improved as denoted by the downwards trend of each of the box plots (Fig. 5.15C). The Interface Quality (IntQual) was the scale that improved the most, while the System Usefulness (SysUse) remained similar. The rest of the scales improved slightly even with the additional features for Phase 2.

Qualitative

The qualitative results included the transcriptions of the think aloud comments of the participants and notes taken by the usability moderator during the experimental sessions. The transcripts and notes are made available in the GitHub repository of this PhD thesis:

https://github.com/navarral/phd-thesis/ (evaluation/phase-2/)

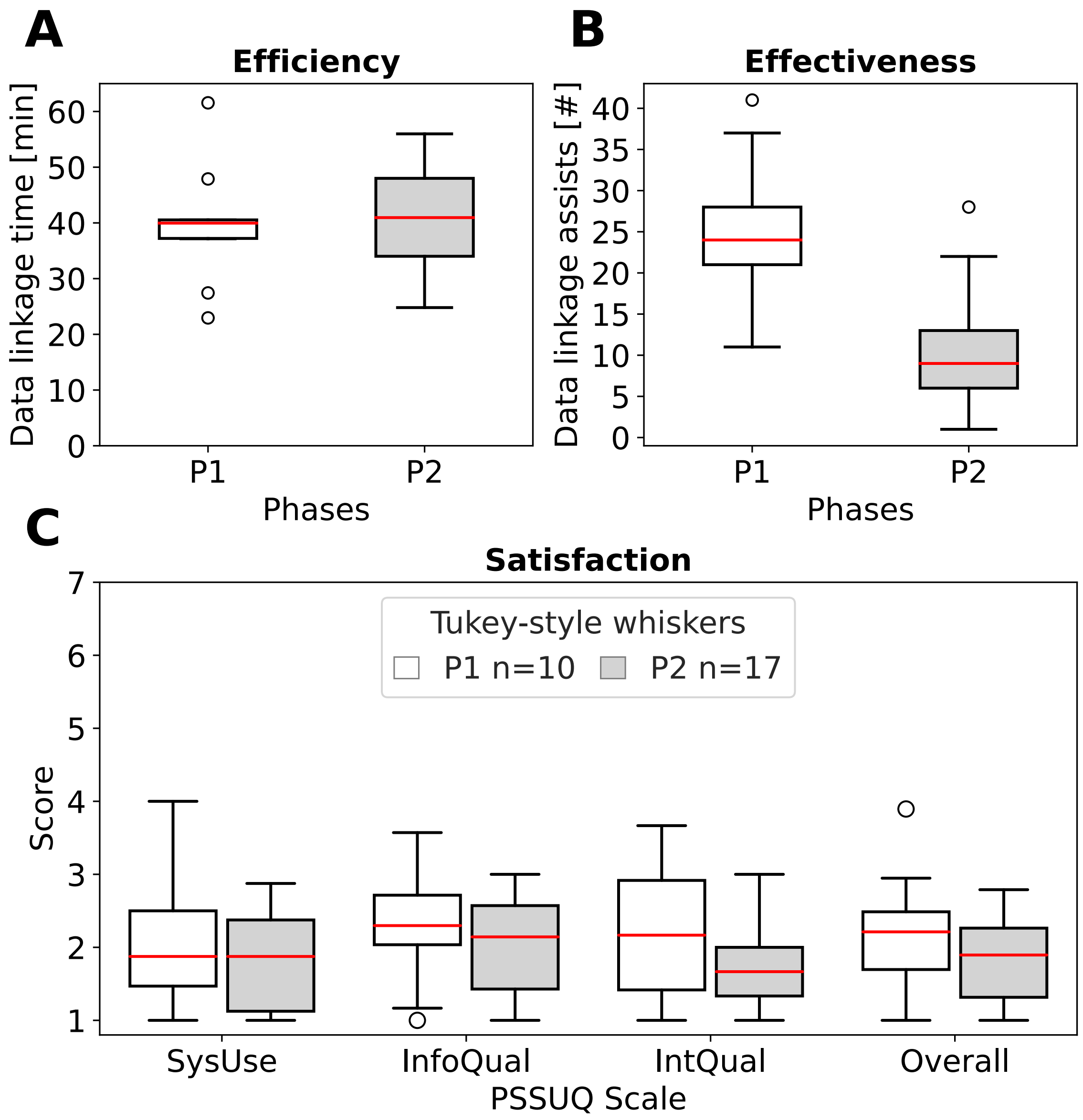


Figure 5.15: Progression of the comparable quantitative metrics for usability collected in the **Phase 1 and 2** of the evaluation (P1, white, n=10; P2, grey, n=17). **A – Efficiency.** Time spent per participant during the data linkage tasks as box plots (tasks T1-6 for P1 and T1-8 for P2), the lower the value the higher the efficiency. **B – Effectiveness.** Number of assists from the moderator during the data linkage tasks as box plots (tasks T1-6 for P1 and T1-8 for P2), the lower the value the higher the effectiveness. **C – Satisfaction.** Comparative PSSUQ scales box plots for System Usefulness (SysUse), Information Quality (InfoQual), Interface Quality (IntQual) and Overall for the P1 and P2 phases of the usability testing. The sample sizes for each of the metrics are 80, 70, 30 and 190 (P1) and 136, 119, 51 and 323 (P2) respectively. The scores are in a Likert 7 points scale where the lower the value the higher the satisfaction.

5.3.4.5 Thematic Analysis

The text transcriptions from the usability sessions were analysed following the six step process of thematic analysis previously outlined in Section 5.1.5. The results of the analysis for Phase 2 are presented as in Phase 1 (Section 5.2.4.4).

Themes and findings summary. The themes name, description, number of references, progress towards achieving the requirements and scope of the finding in terms of the framework artefacts (KG, Methodology and UI) from the participants are summarised in Table 5.7. The analysed themes captured an indication of the progress towards achieving the user requirements important for the overall research question, while also including emerging requirements and the experience when completing the tasks using the user interface.

Table 5.7: Thematic analysis summary of the usability sessions transcripts colour coded as in Fig. 5.17 heat map.

Themes	Findings	References	Scope
Requirement 1: Querying	While querying environmental data associated with particular events was possible, the event concept and approach were complex, confusing the query process.	386	Framework
Requirement 2: Understanding	The metadata is enough to understand the provenance and lineage of the linked data and the visualisations are useful to explore the data but the content is complex and hard to navigate making it not user friendly.	660	UI
Requirement 3: Querying	Exporting the (meta)data is simple and useful to be used as input for analysis.	201	Framework
Emerging Requirements	Additional features and explanations together with simpler words would increase the usability of the framework.	260	Framework
Usability Testing	Overall positive experience when using the dashboard but moderator interventions were needed due to system technical issues and task design.	423	UI, task design

Types of moderator's assistance. The total number of moderator's assists for Phase 2 have been categorised and divided per tasks as in Phase 1 (Fig. 5.16). The discussion of the moderator's task is included as part of the Findings section below.

Codes of the thematic analysis. The resulting codes and themes for Phase 2 are presented as an annotated heat map for traceability as in Phase 1 (Fig. 5.17). The descriptions of the codes for this phase were made available in Appendix D. The discussion of the resulting codes from the thematic analysis is included as part of the *Findings* section below.

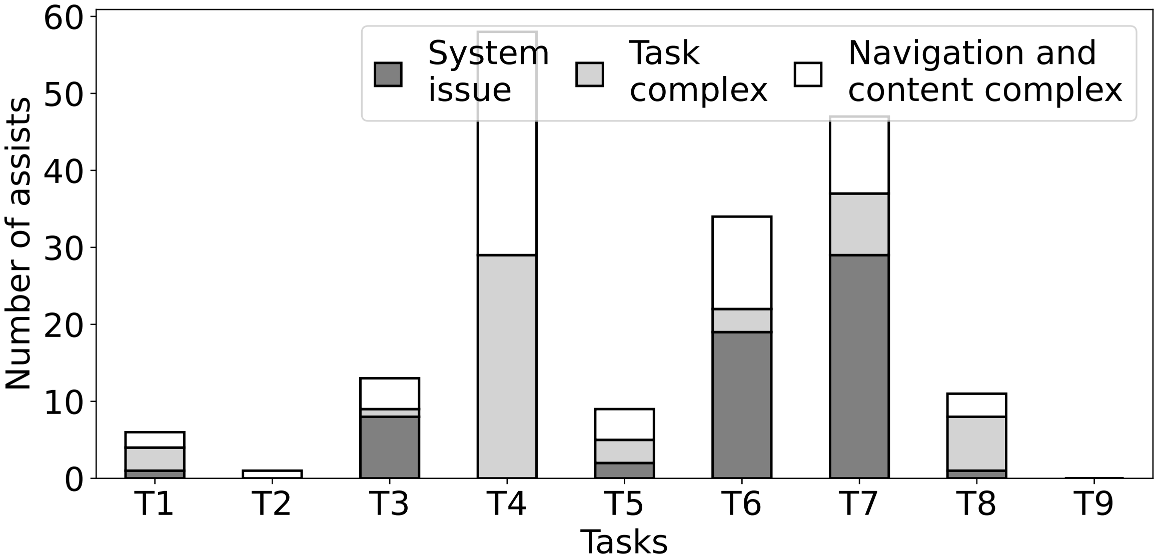
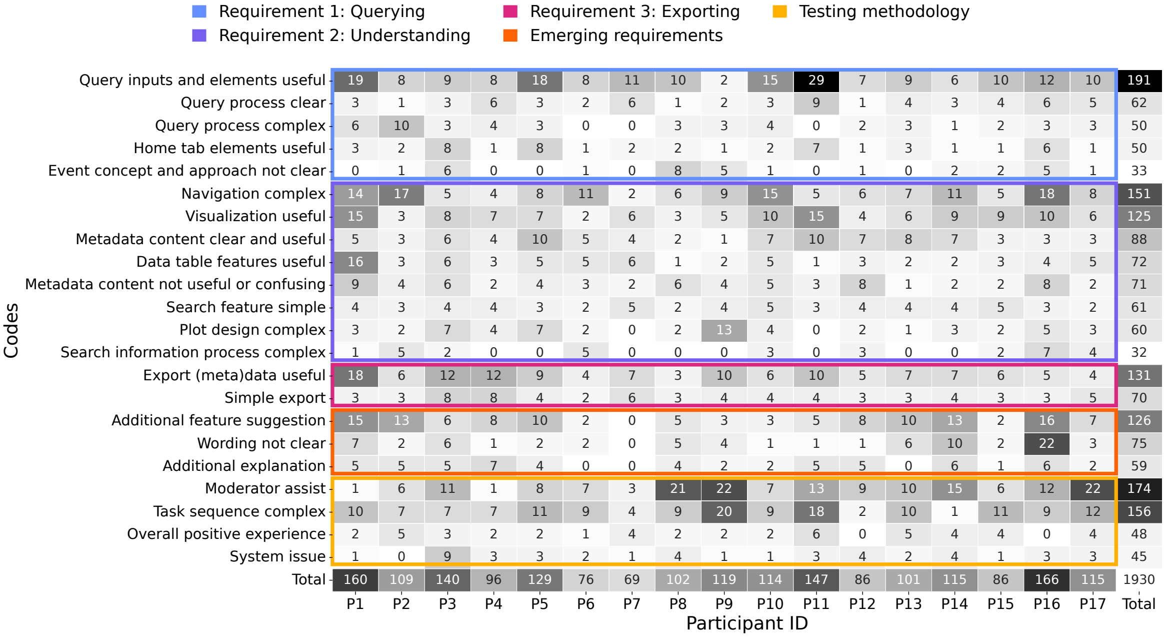


Figure 5.16: Stacked bar plot with number of assists from the usability testing moderator to the participants (n=17) in Phase 2. The type of assistance is included as: the system is crashed or is not responsive (System issue, dark grey), the task is confusing or not completed in a sequential manner (Task complex, light grey) and the design and/or content of the user interface complicates the completion of the tasks (Navigation and content complex, white).



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	Total
Query inputs and elements useful	19	8	9	8	18	8	11	10	2	15	29	7	9	6	10	12	10	191
Query process clear	3	1	3	6	3	2	6	1	2	3	9	1	4	3	4	6	5	62
Query process complex	6	10	3	4	3	0	0	3	3	4	0	2	3	1	2	3	3	50
Home tab elements useful	3	2	8	1	8	1	2	2	1	2	7	1	3	1	1	6	1	50
Event concept and approach not clear	0	1	6	0	0	1	0	8	5	1	0	1	0	2	2	5	1	33
Navigation complex	14	17	5	4	8	11	2	6	9	15	5	6	7	11	5	18	8	151
Visualization useful	15	3	8	7	7	2	6	3	5	10	15	4	6	9	9	10	6	125
Metadata content clear and useful	5	3	6	4	10	5	4	2	1	7	10	7	8	7	3	3	3	88
Data table features useful	16	3	6	3	5	5	6	1	2	5	1	3	2	2	3	4	5	72
Metadata content not useful or confusing	9	4	6	2	4	3	2	6	4	5	3	8	1	2	2	8	2	71
Search feature simple	4	3	4	4	3	2	5	2	4	5	3	4	4	4	5	3	2	61
Plot design complex	3	2	7	4	7	2	0	2	13	4	0	2	1	3	2	5	3	60
Search information process complex	1	5	2	0	0	5	0	0	0	3	0	3	0	0	2	7	4	32
Export (meta)data useful	18	6	12	12	9	4	7	3	10	6	10	5	7	7	6	5	4	131
Simple export	3	3	8	8	4	2	6	3	4	4	4	3	3	4	3	3	5	70
Additional feature suggestion	15	13	6	8	10	2	0	5	3	3	5	8	10	13	2	16	7	126
Wording not clear	7	2	6	1	2	2	0	5	4	1	1	1	6	10	2	22	3	75
Additional explanation	5	5	5	7	4	0	0	4	2	2	5	5	0	6	1	6	2	59
Moderator assist	1	6	11	1	8	7	3	21	22	7	13	9	10	15	6	12	22	174
Task sequence complex	10	7	7	7	11	9	4	9	20	9	18	2	10	1	11	9	12	156
Overall positive experience	2	5	3	2	2	1	4	2	2	2	6	0	5	4	4	0	4	48
System issue	1	0	9	3	3	2	1	4	1	1	3	4	2	4	1	3	3	45
Total	160	109	140	96	129	76	69	102	119	114	147	86	101	115	86	166	115	1930

Figure 5.17: Codes and themes that emerged from the thematic analysis of the usability sessions transcripts in Phase 2 as a heat map for traceability.

Findings. The author of this thesis observational findings from the Phase 2 usability test (thematic analysis supported by the quantitative metrics) are described in this section together with the recommendations for the next phase.

Theme: Requirement 1: Querying

Finding (P2.F1). *"While querying environmental data associated with particular events was possible, the event concept and approach were complex, confusing the query process."*

Evidence. The 17 expert participants were able to complete the first three tasks (T1, T2 and T3), which were formulated based on the querying requirement (Fig. 5.12). The time per task variability was below 2 minutes for the first three tasks (Fig. 5.13), supporting that querying was possible for all participants in a similar manner. The majority of the code references for the first theme indicate clarity and usefulness of the query inputs, elements and processes (253 out of 386, Fig. 5.17), and the codes were referenced across all participants.

However, the tasks that involved actual querying (T3 and T7) required the most assistance due to system issues related to a memory exceeding issue due to the hosting service used during the usability session. The memory limit was noticed three times more in T7 since the free memory was even less than at the start (T3). The system issue is also reflected in the high PSSUQ score in Q9 (Fig. 5.14).

Furthermore, 14 out of 17 participants found the query process and query inputs complex at least once (Fig. 5.17). That was mostly due to the time window length and lag and the event query inputs, which together made the query process less straightforward. The PSSUQ open comments reinforced the complexity of the query process and the understanding of the event approach (10 out of the 13 comments, Appendix E). While it was understood that environmental data was being gathered for events that happen within a spatial region, the temporal part of the event was less clear. Our hypothesis was that the events exemplified a particular use case but the researchers were thinking of answering their own research question. Therefore, limiting the events to a particular use case may be the source of the confusion in the query process together with the terminology of the time window parameters.

Recommendations:

- P2.F1.1. Host the dashboard in a site, service or virtual machine where memory capacity should not be a problem for a regular query.
- P2.F1.2. Provide an example to visualise the time window parameters and the outcome of the query.

– P2.F1.3. Allow users to input their own events to answer their research questions instead of limiting to the available example event data in the triplestore.

Theme: Requirement 2: Understanding

Finding (P2.F2). *"The metadata is enough to understand the provenance and lineage of the linked data and the visualisations are useful to explore the data but the content is complex and hard to navigate making it not user friendly."*

Evidence. Most of the expert participants needed assistance when completing T4, T5, T6 and T7, which are tasks formulated towards the understanding of the event-environmental linked data (Fig. 5.12). The task completion pattern together with the time per task magnitude and variability of tasks T4, T5, T6 and T7 indicate that these were complicated tasks compared to the rest (Fig. 5.12 and 5.13). The results are also coherent with the number of subtasks within each task.

Furthermore, the most complicated task, T4, was associated with understanding where data came from (i.e. data provenance) and the processing steps from the initial data sources to the aggregated version (i.e. data lineage). Therefore, learning about the data from understanding the content of the metadata. While the metadata was useful to learn about the data, the content was complex and searching for the information was not straightforward. This statement is supported by the codes emerged regarding the metadata from the usability transcripts (Fig. 5.17) and PSSUQ open comments (Appendix E) together with the Q12 PSSUQ score (Fig. 5.14).

The visualisation of the data as a data table and with plots were useful for the participants. The features in the data table were appreciated when trying to have a first insight of the queried data. While the plots were also useful to explore the data from different perspectives, they needed an additional explanation to clarify some of the elements and guide the participants in what they should be looking for (Fig. 5.17).

Nevertheless, the most negative finding of the usability study was that the navigation was complex as supported by the type of assists for tasks T4, T5, T6 and T7 (Fig. 5.16), and the code with the most references from the Requirement 2: Understanding theme in the usability transcripts (151 out of 660, Fig. 5.17) and the PSSUQ open comments (12 out of 28, Appendix E). Therefore, the layout of the dashboard needs to be improved to be more inline with the workflows and story lines from the researchers.

Recommendations:

– P2.F2.1. Metadata content needs to be simplified and summarised in a user friendly way.

- P2.F2.2. Visualisations need explanations to guide researchers in the interpretation of the results.

- P2.F2.3. The dashboard approach with multiple tabs will be switched to a simplified user interface with 3 main sequential steps: (i) upload the event data, (ii) select query options and (iii) exportable and explorable output.

Theme: Requirement 3: Exporting

Finding (P2.F3). *"Exporting the (meta)data is simple and useful to be used as input for analysis."*

Evidence. All of the expert participants completed T8 (Fig. 5.13), which was related to exporting the data and metadata. However, 8 participants needed a one time intervention due to the complexity of the task (5) or because the navigation was complex (3) (Fig. 5.5.16). Only one participant needed 3 interventions, 2 related to the task complexity and 1 to the complex navigation (Fig. 5.16). The short time spent in T8 and the low variability below 2 min support the simplicity of the (meta)data export (Fig. 5.13).

The number of references for the codes related to the usefulness (131) and simplicity (70) to export the (meta)data (Fig. 5.17). Exporting the data for a subsequent analysis was imperative for researchers. The metadata was useful to understand and contextualise the data for the researcher. This allows other researchers to reuse the data after publication. However, some of the comments included in the navigation complex code referred to the confusion around the graph formats (i.e. RDF and TTL) of the exports (Fig. 5.17).

Recommendations:

- P2.F3.1. Keep the export button to download the linked data and metadata

- P2.F3.2. Facilitate the understanding of the graph formats with a human-understandable output.

Theme: Emerging Requirements

Finding (P2.F4). *"Additional features and explanations together with simpler words would increase the usability of the framework."*

Evidence. The overall framework implementation was useful for the participants, which is represented by the Q19 and SysUse scores, but the framework needs additional features denoted by the higher score in Q18 (Fig. 5.14). Most of the references for the forth theme, emerging requirements, were related towards extending the current features of the framework in the usability transcripts (126 out of 260, Fig. 5.17) and PSSUQ open comments (21 out of 29, Appendix E) thematic analysis. All participants but one suggested additional features, which would increase the usability of the framework. The additional features included: (i) adding an advanced aggregation method and selecting all query features, (ii) summarising the metadata in a user-friendly way, (iii) facilitating the understanding of the data table and increasing the amount of data, (iv) defining, selecting and grouping events feature for exploration, (v) adding a time series, scatter and histogram plots, and (vi) distinguishing outliers based on historical data.

Even though the framework is specialised for health-environmental research, some of the technical terms and paragraph wording were hard to understand. All but one participant referred to the confusing wording at least twice with an overall reference number of 75 out of 260 (Fig. 5.17). In particular, the terms in the metadata content, which required additional explanation. Another element that required further explanation was the plots to explore the data. The number of references in the usability sessions transcript (59 out of 260, Fig. 5.17) and the PSSUQ open comments (7 out of 29, E) support the need for additional explanations to facilitate the use of the framework.

Recommendations:

- P2.F4.1. Include the additional features suggested by the users.
- P2.F4.2. Simplify the wording when possible and include a tooltip providing a definition for complicated terms.
- P2.F4.3: Include an explanation for all the visualisations in the dashboard.

Theme: Testing Methodology

Finding (P2.F5). *"Overall positive experience when using the dashboard but moderator interventions were needed due to system technical issues and task design."*

Evidence. The expert participants had an overall positive experience when using the dashboard to complete the tasks as supported by the number of references in the usability sessions transcript (48 out of 423, Fig. 5.17) and the PSSUQ open comments (16 out of 26, Appendix E). The Overall scale of the PSSUQ displays a distribution of values between 1 and 2 (Fig. 5.15C), which reinforces the positive results from the thematic analysis. The overall and interface quality satisfaction have improved in respect to the P1 usability evaluation (Fig. 5.15C).

However, the participants required assistance to complete most of the tasks related to system issues and the task being complex to understand (Fig. 5.16). The system issues lead to the non completion of T7 for 5 out of 17 participants (Fig. 5.12) due to the memory limit explained in Finding P2.F1. Following, some of the plots did not work for certain environmental variables, mostly for pollutants, leading to confusion.

The participants also needed assistance in understanding the tasks or found the tasks complex (Fig. 5.17), which could mean that the sequence of tasks did represent a real workflow or that the wording of the tasks was not intelligible. In addition, the PSSUQ score distribution for Q4 and Q5, which were related to the quickness and efficiency when completing the tasks, were higher than the rest of the questions (Fig. 5.14). However, the effectiveness of the framework improved from Phase 1 (Fig. 5.15B). Most of the assists due to a task being complex were for T4 (29 out of 54, Fig. 5.16), which represented a particular step in the workflow that the participants had trouble reinforcing the discussion about metadata not being clear in the second finding.

Recommendations:

- P2.F5.1. Check why some of the environmental variables do not display a plot.
- P2.F5.2. The wording and story line of the tasks will be improved and simplified.

5.3.4.6 Conclusions

The second iteration of SERDIF evaluated with **Use Case 2 - KD in Japan** results indicate a promising outcome in that they indicate that the framework is potentially useful in allowing researchers themselves to link health and environmental data whilst hiding the complexities of the use of KG (P2.F1).

The summative usability part of the evaluation granted the quantitative data to compare effectiveness, efficiency and satisfaction against the benchmarks set in Phase 1. Overall, the usability of the framework has improved in Phase 2 based on the number of moderator's assists when completing the data linkage tasks (effectiveness), and the PSSUQ scales, which improved or remained the same compared to Phase 1. SERDIF usability improved even though new functionalities and features have been added to the framework required by the users.

The formative usability part of the evaluation indicates that Phase 2 was successful in deriving findings and recommendations for the further development of SERDIF. While researchers can link particular health events with environmental data to explore environmental risk factors of rare diseases, SERDIF needs to be refined to enhance the linkage process (P2.F1). In particular, the SERDIF UI component needs to be simplified in terms of its content, navigation and choice of data visualisations (P2.F2); while keeping the exporting functionality (P2.F3). Additional features mentioned by the participants and facilitating the understanding of the technical jargon and visualisations would improve the usability of SERDIF (P2.F4). Regarding the tasks, simplifying the content and improving the task design would also improve the overall usability (P2.F5).

The increase in usability between the two phases and the evidence towards the framework being potentially useful support the adequacy of the usability evaluation approach. The increase in the sample size between P1 and P2 together with the positive results support the generalisation of the framework beyond a single case study.

5.4 Usability Test - Phase 3

The third usability test of the study is presented in this section. The content of this section is structured following the four steps of a user-centred design as in Phase 1 and 2.

5.4.1 Context of Use

Name. Use Case 3 - AAV in Europe
Description. Same as **Use Case 1 - AAV in Ireland** (Section 5.2.1).
Research projects. The FAIRVASC project was added to the research projects from **Use Case 2 - KD in Japan**. The FAIRVASC project aims to link AAV registries across Europe towards the vision of a 'single European dataset' for vasculitis [FAIRVASC, 2022]. The consortium has already overcome the challenge of harmonising the different datasets by uplifting the health data to RDF in each of the registries and

then running SPARQL queries to link the datasets. However, HDRs have not yet linked environmental data with the existing clinical data.

This use case did not only test the scalability of SERDIF but it generalised once more the usefulness of the data linkage approach across researchers around Europe.

5.4.2 Expert User Requirements

The expert user requirements were refined with minor specifications after the analysis of the evaluation results in Phase 2. The refinement included (R1.2) a flexible input of health events to make them relevant to each researcher's studies (P2.F1.3); and (R2.2) a more simplified view focused on the data linkage process with additional information on demand (P2.F2.3). The expert user requirements evaluated in Phase 3 are the following with the underlined words denoting the refinements of the requirements:

Requirement 1 (R1.2). *Enable HDRs to query environmental data associated with relevant/own individual health events through location and time, within the area of the event and a period of data before the event.*

Requirement 2 (R2.2). *Support the understanding of event-environmental linked data and metadata, with its use, limitations and data protection risk for individuals, by using a simplified view focused on the data linkage process with optional further information.*

Requirement 3 (R3.2). *Export event-environmental linked (meta)data to be used as input in statistical models for data analysis (CSV) and for publication (RDF).*

5.4.3 Framework Implementation

The SERDIF components have been updated after the results from Phase 2 of the usability study and the refined requirements (Section 5.3.4). The updated components are the Methodology, KG and UI components of the framework. The KG and UI were moved to one of the virtual machines reserved for research projects at the ADAPT centre in Trinity College Dublin (P2.F1.1).

5.4.3.1 Methodology Component

The *Step 4: Data visualisation* and *Step 5: Data export/Downlift* of the SERDIF methodology were simplified into a unique step, Step 4: Data interaction. The interaction with the KG consists of a UI to facilitate data linkage process and an export of the linked data and metadata generated as a data table for analysis, a graph for publication and an interactive report for exploration.

5.4.3.2 Knowledge Graph Component

The requirement of researchers to be able to query health events relevant to their research resulted in an update of the SERDIF KG component after Phase 2 (R1.2 in Section 5.3.4).

5.4.3.3 User Interface Component

Step 1: Data collection (implementation). The weather data source was replaced with the E-OBS daily gridded meteorological dataset for Europe from Copernicus [Copernicus, 2020] after consulting with experts in environmental science (P2.F4.1). The experts agreed on this particular dataset because the grid points have been computed from in-situ observations, which are more appropriate than other data sources (e.g. reanalysis data from satellites) for this particular case study. Regarding the coverage, the dataset covers the surface level for all European countries (spatial) and a subset for 2011-2020 was selected (temporal) since the virtual machine provided from the ADAPT centre in Trinity College Dublin had a data storage limit (P2.F1.1). The air pollution data from EEA was expanded to include coverage for the following countries besides Ireland: Czech Republic, Switzerland and United Kingdom (P2.F4.1). The selection of the countries was based on the area of interest from the participants recruited for this use case.

The NUTS (Nomenclature of territorial units for statistics classification) geometry data representing the regions in the EU [eurostat, 2021] was also included to enable the spatial linkage required.

The example health events data was removed from the KG. The health context for the environmental data will now be included at the SPARQL query level (*Step 3: Data linkage (implementation)* in this Section). The only data types collected in this step are weather, air pollution and geometry data, providing the researchers with the flexibility requested in the input of health events data (P2.F1.3).

Step 2: Semantic uplift (implementation). The weather and air pollution data were uplifted to RDF following the same QB description as in Phase 2. In this case, the weather grid points are described as individual datasets where the location is specified at the dataset metadata level (Listing 5.5) and the observations are modelled as time series following a multi-measure approach [Cyganiak and Reynolds, 2014] (Listing 5.6 bottom part and [Navarro-Gallinad, 2021] for the complete RDF graph file). The total number of triples for the environmental data increased from 80M to 470M between Phase 1 and Phase 3.

The NUTS geometry data representing the regions in the EU [eurostat, 2021] was converted from NeoGeo to GeoSPARQL structures using a Construct SPARQL query [Navarro-Gallinad, 2022]. The purpose of the conversion was to enable GeoSPARQL spatial reasoning features for geometry-based queries.

The R2RML mappings and ontologies used in Phase 3 are made available in the GitHub for the thesis to reproduce the uplift process:

https://github.com/navarral/phd-thesis/ (implementation/)

Step 3: Data linkage (implementation). The Construct SPARQL query to link the health and environmental data was edited to include the user's input health events data (Listing 5.7). The other query elements remained the same from Phase 2 (Listing 5.2 and Listing 5.4).

Listing 5.7: Snippet of the additional query elements to include the user's input health context.

```
# -- Namespaces -------------------------------------
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geosparql: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX locn: <http://www.w3.org/ns/locn#>
# -- Include user's event ID ------------------------
BIND(IRI("https://serdif.adaptcentre.ie/kg/2022/event#event-42")
     AS ?eventRef)
# -- Select time window from user's input -----------
BIND(xsd:dateTime("2015-12-01T00:00:00Z") AS ?evDateT)
BIND(?evDateT - "P0D"^^xsd:duration AS ?dateLag)
BIND(?dateLag - "P90D"^^xsd:duration AS ?dateStart)
# -- Select datasets within an area from user's input ---
BIND( "POINT(8.549 47.366)"^^geosparql:wktLiteral AS ?point)
?area a geosparql:Feature ;
     rdfs:label ?areaName ;
     geosparql:hasGeometry/geosparql:asWKT ?geo ;
                        ramon:level 3 .
FILTER(geof:sfWithin(?point, ?geo))
GRAPH serdif:metadata {
    ?qbDataSet a qb:DataSet, geosparql:Feature ;
             locn:geometry ?qbGeoB .
    ?qbGeoB geo:asWKT ?qbGeo.
}
FILTER(geof:sfWithin(?qbGeo, ?geo))
```

In addition, two more queries were added in the linkage process to construct datasets describing the environmental context for the area and season (or month) of the event [Navarro-Gallinad et al., 2023] (Listing 5.8). The environmental context refers to what values are historically normal to have during a particular month in a particular region (P2.F4.1). For example, a temperature of 20ºC might be considered a high temperature for an event that happened in Dublin during the month of October, but an average temperature for an event in Barcelona for the same month. Without this environmental context, researchers would need additional knowledge from experience or compute the historical averages to be able to study and interpret events that took place in different areas and seasons (P2.F2.2).

Listing 5.8: Snippet of the additional query elements to include the environmental context for the data.

```
# -- Namespaces ----------------------------------------------------
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ofn: <http://www.ontotext.com/sparql/functions/>
# -- Historical average monthly value (envVarN) --------------------
SELECT ?event ?monthN ?envPropN (AVG(?envVar) AS ?envVarN)
BIND(MONTH(?obsTime) AS ?monthN)
# -- Historical standard deviation monthly value (sdev) ------------
SELECT ?event ?monthN ?envPropC (xsd:float(ofn:sqrt(SUM(?envVarNmean)/(?
    envVarC - 1))) AS ?sdev)
SELECT ?monthN ?envProp ?envVarC ((?envVar - ?envVarN)*(?envVar - ?envVarN)
    AS ?envVarNmean)
```

The SPARQL queries used in Phase 3 are made available in the GitHub for the thesis to reproduce data linkage process:

https://github.com/navarral/phd-thesis/ (implementation/)

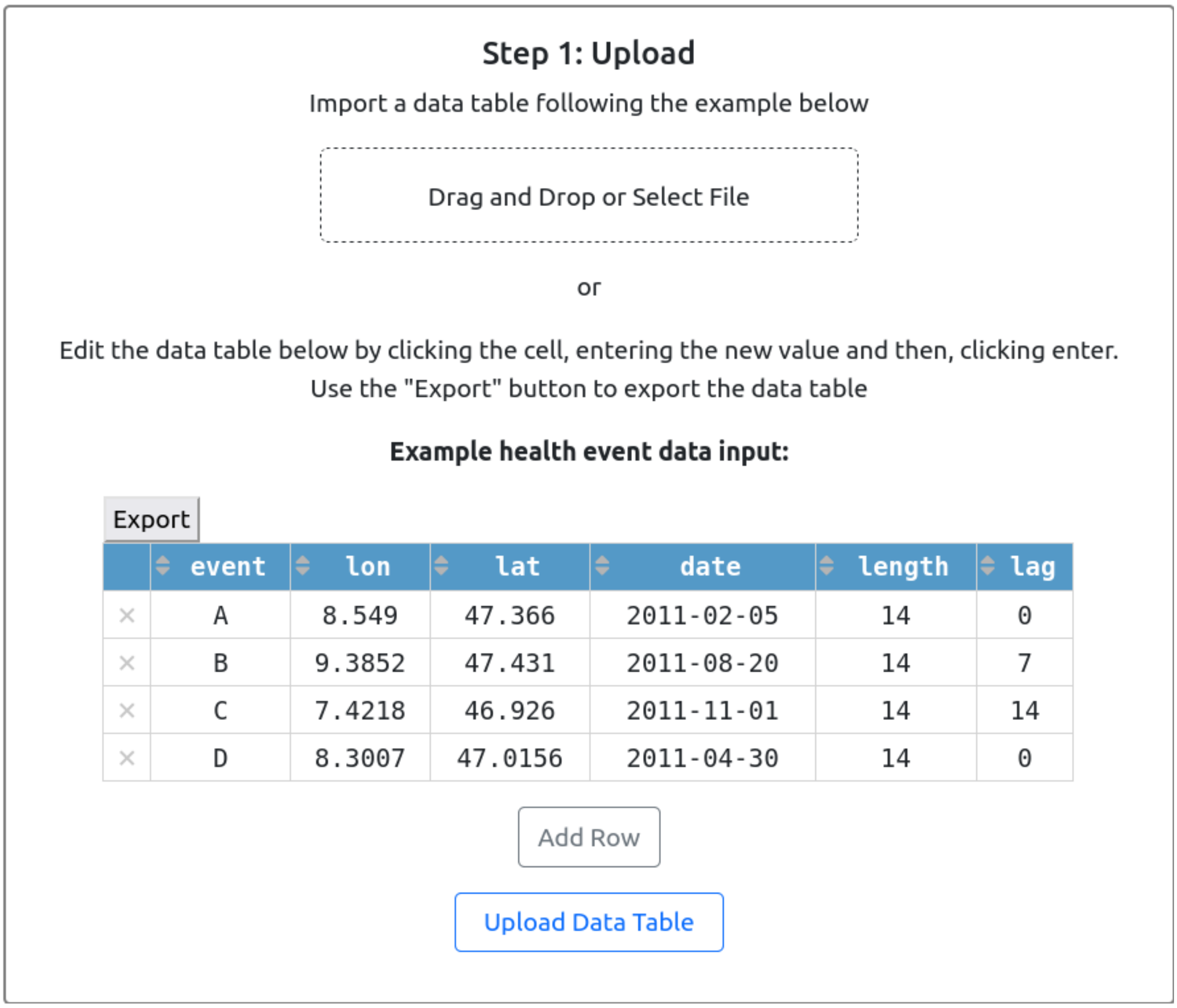
The implementation of *Step 4: Data visualisation* and *Step 5: Data export/downlift* was updated to address the researchers requirements of Phase 3 towards a UI with a simplified view with the focus on the data linkage process (Section 5.4.2) (P2.F2.3).

Step 4: Data interaction (implementation). The data visualisation step implementation was separated into a UI for the linkage process and an interactive report to explore the resulting linked data (P2.F3.2). For a better understanding of the data visualisation step, the UI and interactive report were made available at:

https://w3id.org/serdif/

The 3-step UI. The UI was updated to present the data linkage process in 3 steps: (i) upload the event data, (ii) select query options and (iii) exportable and explorable output (Fig. 5.18 and 5.19) (P2.F2.3 and P2.F3.2).

In the upload step, researchers can input health events relevant to their research by editing an example data table provided or importing a CSV table following the same format (Fig. 5.18) (P2.F1.3). The data table contains the minimum information to link the health events with environmental data: event identifier (event), lat/lon coordinates (lat, lon), date of the event (date), and the time window parameters to gather environmental data relative to the date of the event (length, lag). In addition, a diagram representing the linkage process, including the spatial and temporal components, and an example output were included in an introductory paragraph before the first step of the UI (see the UI URL for Phase 3) (P2.F1.2).



Step 1: Upload

Import a data table following the example below

or

Edit the data table below by clicking the cell, entering the new value and then, clicking enter.
Use the "Export" button to export the data table

Example health event data input:

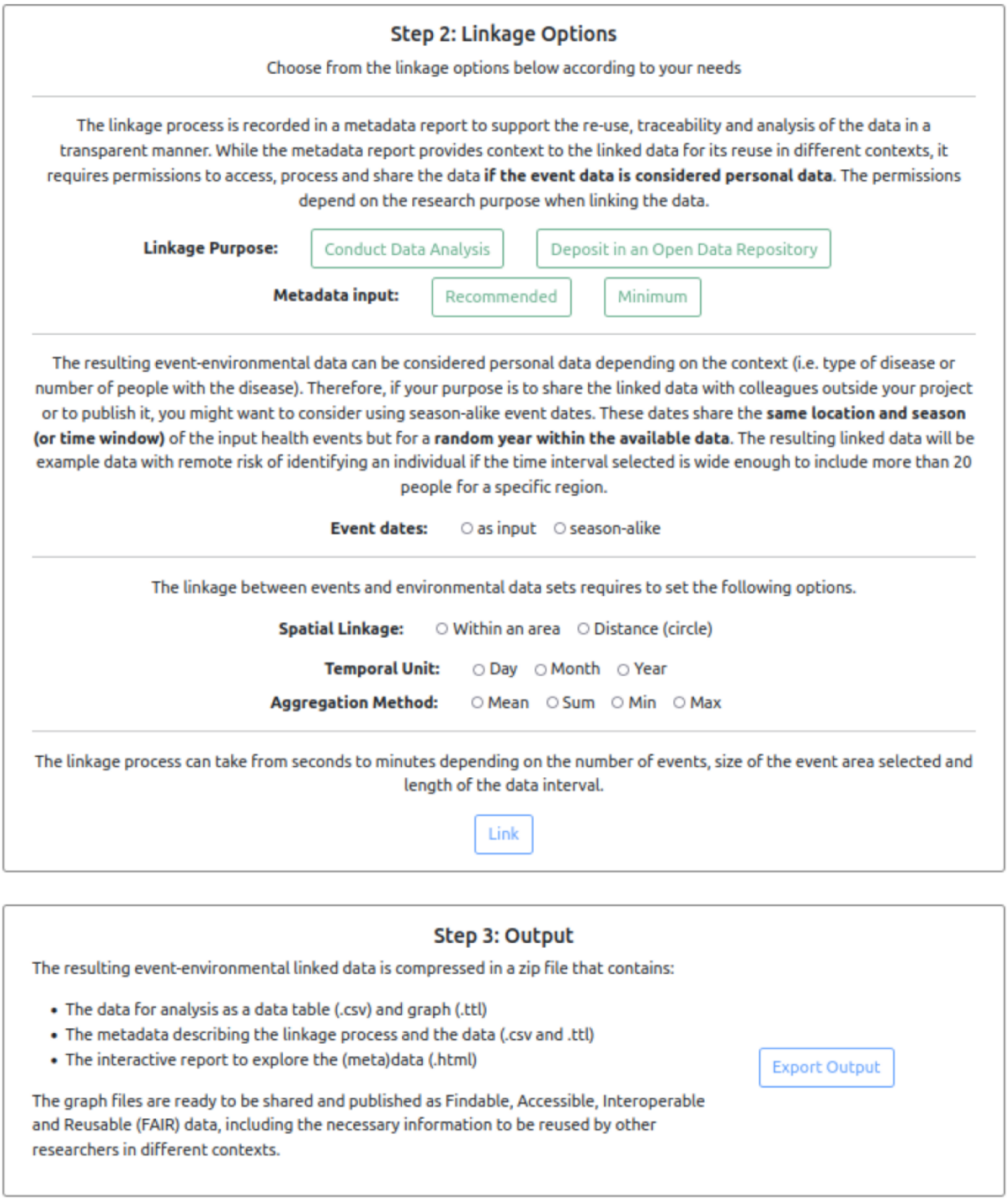
	event	lon	lat	date	length	lag
×	A	8.549	47.366	2011-02-05	14	0
×	B	9.3852	47.431	2011-08-20	14	7
×	C	7.4218	46.926	2011-11-01	14	14
×	D	8.3007	47.0156	2011-04-30	14	0

Figure 5.18: Screenshot of the SERDIF UI displaying the step to upload the health events data (Step 1:Upload).

In the linkage step, First, the user was reminded of the importance of the data protection aspects based on the type of linkage purpose (Linkage purpose in Fig. 5.19). The user (i.e. the data processor) requires explicit permission from the data controller and/or Data Protection Officer (DPO) assigned to the event data to access, use and process the data, if the event data is considered personal data. Second, the extent of the metadata associated with the health events could be chosen between *Recommended* and *Minimum* based on the user's purpose and needs (*Metadata input* in Fig. 5.19). The metadata information displayed when one of the input options was presented as an editable and exportable data table like in the upload step (Fig. 5.18) (P2.F2.1). Third, an option to generate season-alike dates was made available in the case the purpose of the researcher was to to share the linked data with colleagues outside your project or to publish it (*Event dates input* in Fig. 5.19). Season-alike events share the same location and season (or time window) of the input health events but for a random year within the available data. The resulting linked data would be example data with remote risk of identifying an individual if the time interval selected is wide enough to include more than an acceptable number of people for a specific area, specified by the data controller and DPO. Fourth, an area based on the event point location could be defined to select the datasets within that area for a particular event (*Spatial linkage* in Fig. 5.19). The area options were to use the NUTS territories (*Step 1: Data collection* in this section) or draw a circle around the event (e.g. 20km). The selected datasets within the area can be aggregated at a certain time unit from days to years (*Temporal unit* in Fig. 5.19) and using a specific aggregation method to integrate the datasets (*Aggregation method* in Fig. 5.19). In general in this step, the wording was simplified and tooltips were added to facilitate the understanding of complex technical terms (P2.F4.2).

In the export step, researchers can export a zip file that contains linked health-environmental data for analysis as a data table (CSV) and graph (TTL), the metadata describing the linkage process and the data (CSV and TTL) and an interactive report to explore the (meta)data (HTML) (see [Navarro-Gallinad et al., 2023] for an output example) (P2.F3.1). The graph file distribution of the linked data is an interoperable version suitable for machines to understand but hard to use by domain experts (*Challenge 2: KG usability*). The CSV file distributions for the data and metadata were generated as exports towards addressing this challenge.

Interactive report. The SERDIF report provided an initial exploration of the environmental data linked with the input health events from the SERDIF UI. The aim of the report was to make the linked data easier to understand rather than if you looked at the raw data in knowledge graph or data table formats (P2.F3.2). The report is structured in four sections: (i) introduction, (ii) metadata, (iii) data and (iv) contact information. The metadata section describes the information associated with the linked data including dataset descriptors, input data, processing steps, dataset sharing and dataset structure. Each of these categories summarises the information with icons, tables and plots to make the information easier to understand. The data section presents the link data as an interactive table with sort and filter functions while highlighting



Step 2: Linkage Options

Choose from the linkage options below according to your needs

The linkage process is recorded in a metadata report to support the re-use, traceability and analysis of the data in a transparent manner. While the metadata report provides context to the linked data for its reuse in different contexts, it requires permissions to access, process and share the data **if the event data is considered personal data**. The permissions depend on the research purpose when linking the data.

Linkage Purpose: Conduct Data Analysis | Deposit in an Open Data Repository

Metadata input: Recommended | Minimum

The resulting event-environmental data can be considered personal data depending on the context (i.e. type of disease or number of people with the disease). Therefore, if your purpose is to share the linked data with colleagues outside your project or to publish it, you might want to consider using season-alike event dates. These dates share the **same location and season (or time window)** of the input health events but for a **random year within the available data**. The resulting linked data will be example data with remote risk of identifying an individual if the time interval selected is wide enough to include more than 20 people for a specific region.

Event dates: ○ as input ○ season-alike

The linkage between events and environmental data sets requires to set the following options.

Spatial Linkage: ○ Within an area ○ Distance (circle)

Temporal Unit: ○ Day ○ Month ○ Year

Aggregation Method: ○ Mean ○ Sum ○ Min ○ Max

The linkage process can take from seconds to minutes depending on the number of events, size of the event area selected and length of the data interval.

Link

Step 3: Output

The resulting event-environmental linked data is compressed in a zip file that contains:

- The data for analysis as a data table (.csv) and graph (.ttl)
- The metadata describing the linkage process and the data (.csv and .ttl)
- The interactive report to explore the (meta)data (.html)

The graph files are ready to be shared and published as Findable, Accessible, Interoperable and Reusable (FAIR) data, including the necessary information to be reused by other researchers in different contexts.

Export Output

Figure 5.19: Screenshot of the SERDIF UI displaying the steps to select the data linkage options (Step 2: Linkage options) and to export the linked data output for analysis and publication (Step 3: Output).

high and low values for each of the events (P2.F2.1 and P2.F4.1). In addition, the linked data can be explored through interactive time series plots that display the air pollutant values as multiple lines and the weather values at the background as coloured stripes (P2.F4.1). Information to help the understanding of the visualisations of the environmental data was also provided in a foldable manner (P2.F2.2 and P2.F4.3). The contact section has the contact information of the report developer for any queries from the users. A permanent URL is made available to access an example of the interactive report [Navarro-Gallinad et al., 2023] and screenshots of the report are included in Appendix F.

5.4.4 Evaluation Against the Requirements

This section includes the usability test sample size, participant tasks, evaluation results, analysis and conclusions for Phase 3.

5.4.4.1 Sample Size

The pool of 17 HDRs from **Use Case 2 - KD in Japan** was increased with 6 more from this use case, giving a total of 23 HDRs for Phase 3. The additional 6 researchers of this use case are international professors, researchers and PhD students with fluent English, who are studying or have an interest in studying the health outcomes associated with environmental factors for AAV at a European level in their research.

The increased pool of 23 expert participants enhances the chance to discover >90% of the usability problems that can happen 10% of the time, compared to the 15% occurrence in Phase 2 [Lewis, 2014].

5.4.4.2 Experimental Setup Update

The participants required some technical issues while executing the queries related to a memory limitation characteristic of the free version of the hosting service. In P3, the UI was hosted in one of the Virtual Machines reserved for research projects at the ADAPT centre in Trinity College Dublin (P2.F1.1). The UI and KG were checked for technical errors by running multiple queries and selecting different data linkage options (P2.F4.1). The UI was made available for the participants under the following URL:

https://w3id.org/serdif/

5.4.4.3 Participants Tasks Update

The participants were provided with scenario information to improve their understanding of the tasks (P2.F4.2):

Scenario. You are a health data researcher that wants to study the environmental risk factors associated with health outcomes. Therefore, you need to link environmental data with particular health events affecting individuals or populations. Then, you will be able to use the linked data as input for your data analysis.

Health data researcher: researchers with a health background and statistical or data analysis experience (e.g. clinicians, health information technicians/managers and epidemiologists), or statisticians and data analysts that are studying health related outcomes.

Health event: examples of health events are the development of a disease or symptoms, an injury, responding to a medicine, a peak in flu cases or hospital admissions in a certain geographical area.

The tasks were rephrased using simpler language and the subtasks were removed to reduce the overload of information asked in the tasks (P2.F4.2). The number of tasks was reduced to 5 from the previous 7 and 9 from P1 and P2, respectively. The tasks for Phase 3 are presented in Table 5.8.

Table 5.8: Tasks that the participants are asked to complete during the usability testing associated with the requirements from Section 5.4.2 in Phase 3.

Task	**Requirement**	**Description**
T1	Querying	Link environmental data to relevant (or example) health events for your research.
T2	Understanding, exporting	Export the data linkage output and explore the interactive report generated (.html).
T3	Summary	Discuss if you are confident in using the linked data for your research.
T4	Emerging requirements	Explain if you would need any additional features or information before starting the analysis of the linked data.
T5	Summary	Summarise verbally your overall experience when linking data using SERDIF.

5.4.4.4 Results

The quantitative results of Phase 3 include task completion (Fig. 5.20), time per task (Fig. 5.21), the PSSUQ scores and scales (Fig. 5.22), the usability (Fig. 5.23) and testing methodology progress (Table 5.9).

Quantitative

Task completion. The tasks were completed successfully by all of the participants in Phase 3 (Fig. 5.20). The participants required assistance from the moderator at least one time in the tasks associated with the data linkage process (T1-2). Most of the participants did not require assistance when discussing how confident they were in using the linked data for their research (T3). The participants did not need assistance when asked to elaborate on additional features that they would need before starting the analysis of the linked data (T4) and to summarise the overall experience in using SERDIF (T5).

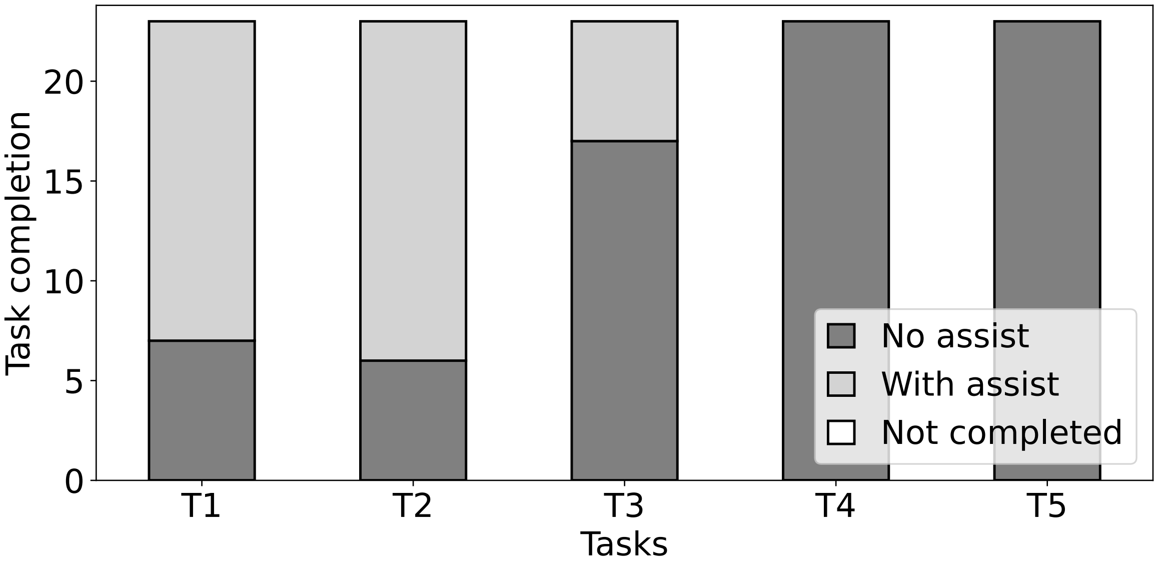


Figure 5.20: Task completion bar plot for Phase 3including the if assistance was needed from the participants (n=23): no assistance required (no assist, dark grey), with at least one intervention to assist (with assist, light grey) and task or subtasks not completed (not completed, white).

Time per task. The box plots in Fig. 5.21 compared participants' time spent on tasks in Phase 3. The time per task increased from T1 to T2 and it dropped significantly for the rest of the tasks T3-5. The tasks associated to data linkage (T1-2) had a larger IQR (∼8min) compared to the tasks where the participant had to discuss the confidence in using the output linked data (T3), explain additional features that may be needed (T4) and summarise their experience in using SERDIF (T5) (∼2min). Overall, the expert participants spent a mean of 28min in the data linkage process (T1-2).

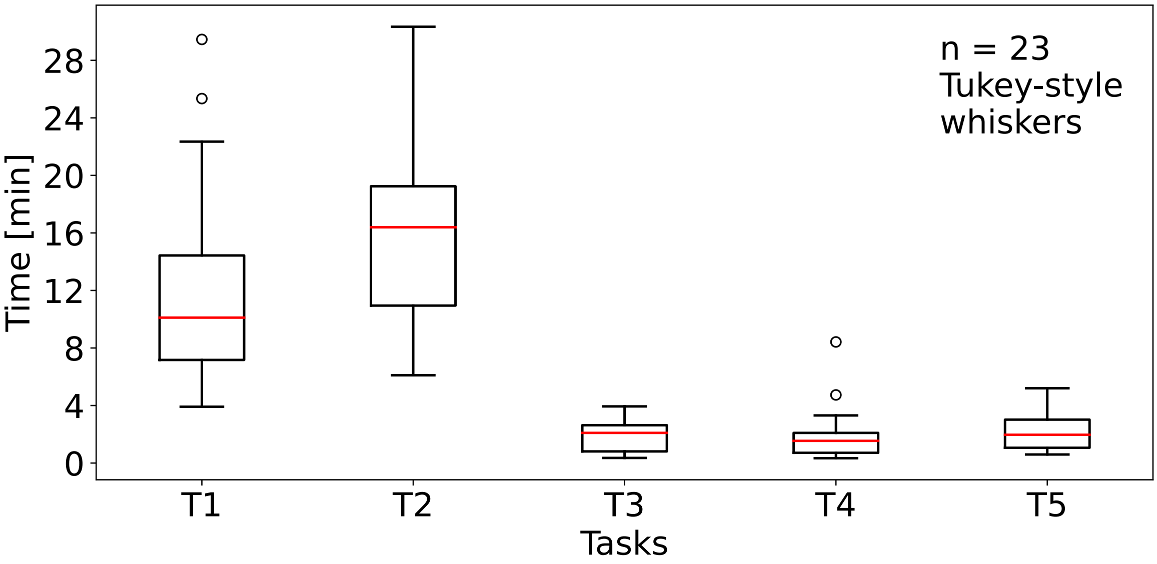


Figure 5.21: Time spent magnitude and variability to complete each task during the usability session represented as a box plot for Phase 3.

PSSUQ scores and scales. All the box plots for the PSSUQ scores but one (Q9) in Fig. 5.22 had a median of between 1 (6 out of 19 questions) and 2 (12 out of 19 questions). Most of them have an IQR below 1 point (14 out of 19 questions). Error messages (Q9) was the worst score per question (2.5); effectively completing the tasks (Q3), becoming productive (Q8), clear information organisation (Q15), pleasant interface (Q16-17), and overall satisfaction (Q19) represent the best scores. The PSSUQ scales with the average scores SysUse, InfoQual, IntQual and Overall had similar values (median ∼1.5) and IQR range (∼0.9) with IntQual being slightly better than the rest. The Overall scale had a median of 1.79, where 1 is the best possible score and 7 the worst.

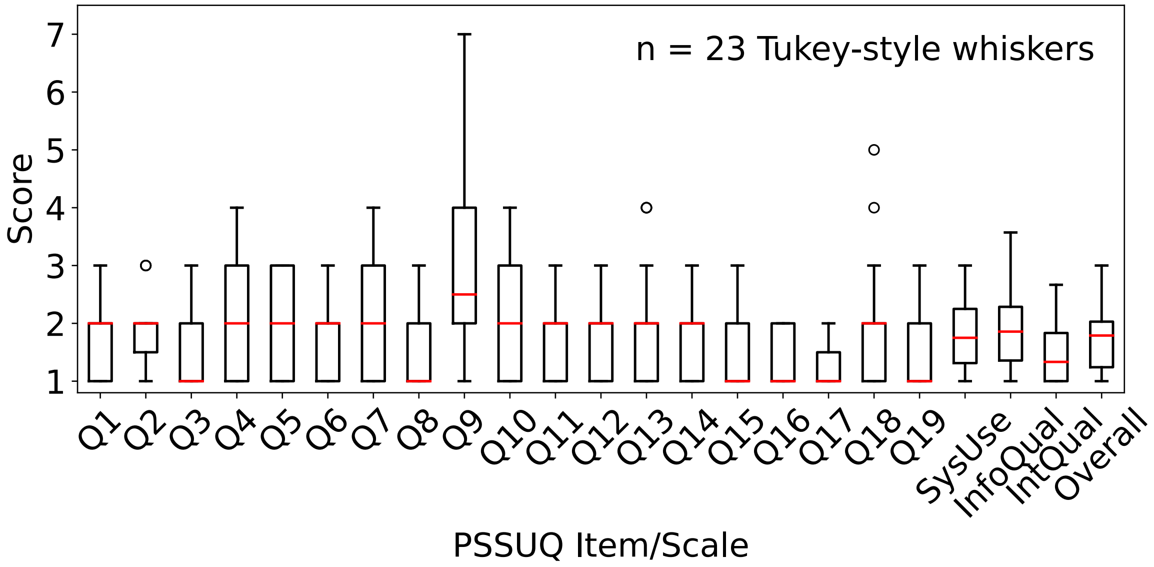


Figure 5.22: PSSUQ scores box plots for Phase 3 with the four averaged metrics (SysUse, InfoQual, IntQual and Overall) on the right end with sample sizes of 184, 161, 69 and 437. The scores are in a Likert 7 points scale where the lower the value the higher the satisfaction.

Usability progress between Phase 1, 2 and 3. The progress of usability metrics (efficiency, effectiveness and satisfaction) across the three phases is presented in Fig. 5.23. The efficiency improved from a median of 40 min to 28 min between Phase 2 and 3 for the data linkage tasks (Fig. 5.23A). The effectiveness improved once more in Phase 3 as denoted by the lower number of assists required per participants when completing the data linkage tasks (Fig. 5.23B). The satisfaction identified from the PSSUQ scales has improved slightly from Phase 2 (Fig. 5.23C). The Interface Quality (IntQual) was the scale that improved the most again, while the System Usefulness (SysUse) remained similar. The Information Quality (InfoQual) and Overall scales improved slightly while the System Usefulness (SysUse) remained similar.

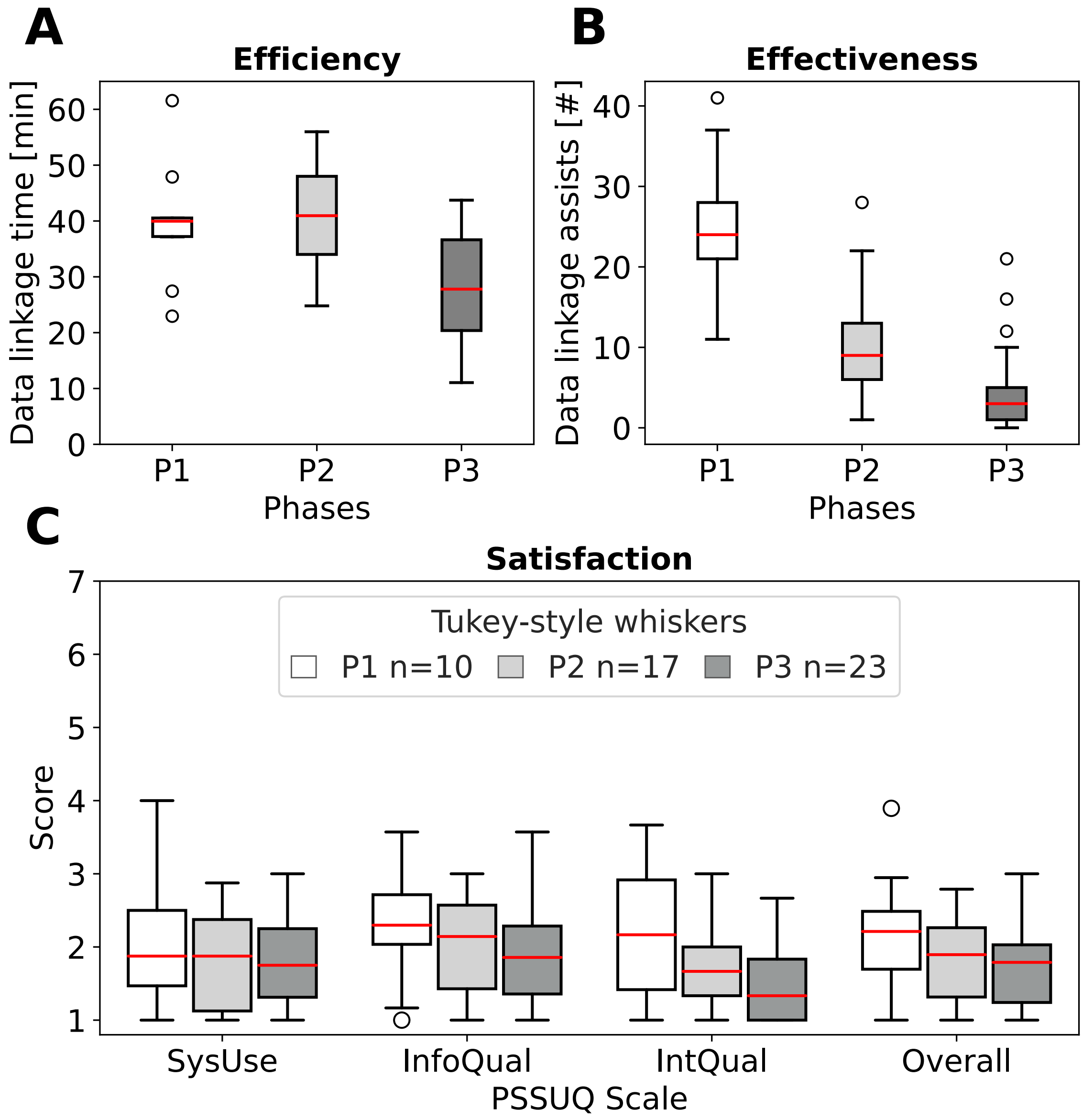


Figure 5.23: Progression of the comparable quantitative metrics for usability collected in the **Phase 1, 2 and 3** of the evaluation (P1, white, n=10; P2, grey, n=17; P3, dark grey, n=23). **A – Efficiency.** Time spent per participant during the data linkage tasks as box plots (tasks T1-6 for P1, T1-8 for P2 and T1-2 for P3), the lower the value the higher the efficiency. **B – Effectiveness.** Number of assists from the moderator during the data linkage tasks as box plots (tasks T1-6 for P1, T1-8 for P2 and T1-2 for P3), the lower the value the higher the effectiveness. **C – Satisfaction.** Comparative PSSUQ scales box plots for System Usefulness (SysUse), Information Quality (InfoQual), Interface Quality (IntQual) and Overall for the P1, P2 and P3 phases of the usability testing. The sample sizes for each of the metrics are 80, 70, 30 and 190 (P1); 136, 119, 51 and 323 (P2); and 184, 161, 69 and 437 (P3) respectively. The scores are in a Likert 7 points scale where the lower the value the higher the satisfaction.

Qualitative

The qualitative results included the transcriptions of the think aloud comments of the participants, the notes taken by the usability moderator during the experimental sessions and the PSSUQ open comments. The transcripts and notes were made available in the GitHub repository of this PhD thesis:

https://github.com/navarral/phd-thesis/ (evaluation/phase-3/)

5.4.4.5 Thematic Analysis

The text transcriptions from the usability sessions were analysed following the six step process of thematic analysis previously outlined in Section 5.1.5. The results of the analysis for Phase 3 are presented below.

Themes and findings summary. The themes name, description and scope of the finding in terms of the framework artefacts (Methodology, KG and UI) from the participants are summarised in Table 5.9. The themes capture the achievement of the expert user requirements and the potential uptake for real use cases, while also identifying some minor improvements for the framework and the need of moderator's guides due to the lack of preparation from the participants for the usability sessions.

Table 5.9: Thematic analysis summary of the usability sessions transcripts colour coded as in Fig. 5.25 heat map.

Themes	Findings	References	Scope
Requirements achieved	HDRs can link health events and environmental data for their research using SERDIF.	796	Framework
Potential uptake	SERDIF can be applied for real use cases with tailored environmental data and features even for non-technical researchers.	189	Framework
Minor improvements	Some important features and text descriptions are not clear in the UI.	429	UI
Testing methodology	The lack of preparation from the participants increased the need for moderator's guidance.	202	Testing

Types of moderator's assistance. The total number of moderator's assists for Phase 3 have been categorised and divided per tasks (Fig. 5.24). The discussion of the moderator's task is included as part of the *Findings* section below.

Codes of the thematic analysis. The resulting codes and themes for Phase 3 are presented as an annotated heat map for traceability (Fig. 5.25). The descriptions of the codes for this phase were made available in Appendix D. The discussion of the codes of the thematic analysis are included as part of the *Findings* section below.

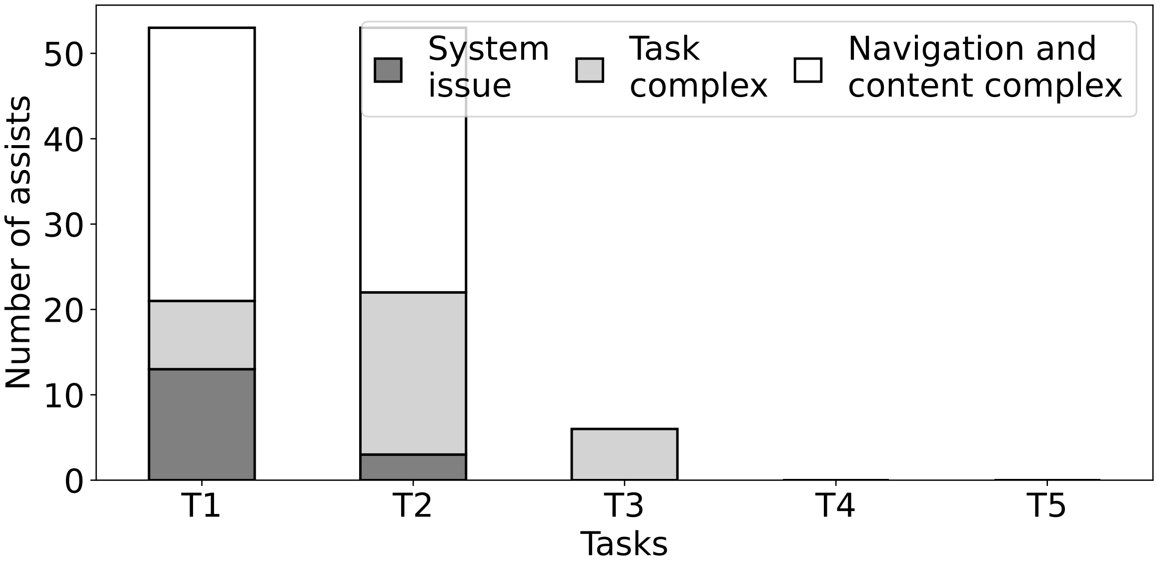


Figure 5.24: Stacked bar plot with number of assists from the usability testing moderator to the participants (n=23) in Phase 3. The type of assistance is included as: the system is crashed or is not responsive (System issue, dark grey), the task is confusing or not completed in a sequential manner (Task complex, light grey) and the design and/or content of the user interface complicates the completion of the tasks (Navigation and content complex, white).

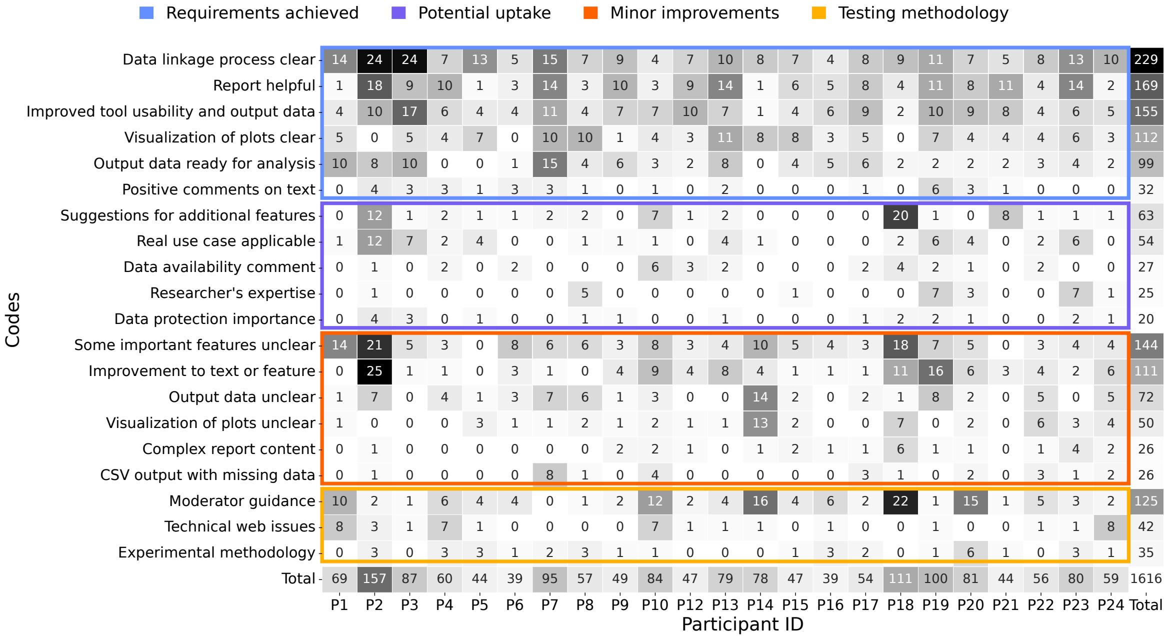


Figure 5.25: Codes and themes that emerged from the thematic analysis of the usability sessions transcripts in Phase 3 as a heat map for traceability.

Findings

The author of this thesis observational findings from the Phase 3 usability test (thematic analysis supported by the quantitative metrics) are described in this section together with the recommendations for the next phase.

Theme: Requirements achieved

Finding (P3.F1). *"Health Data Researchers (HDR) can link health events and environmental data for their research using SERDIF."*

Evidence. The most referenced theme in the third usability test is Requirements achieved (796 out of 1616, Table 5.9 and Fig. 5.25). The references for the codes in this theme were distributed homogeneously indicating an overall consensus with the finding (Fig. 5.25).

The most referenced code identifies the linkage process as clear, from uploading the health data to understanding and exporting the linked environmental data, by the expert participants (229 out of 1616, Fig. 5.25). The rest of the codes in this first theme referred to the improved approach to facilitate the linkage process. The 3-step structure of the UI (1-upload, 2-link and 3-export) increased the overall usability of the tool and the data (155 out of 796, Fig. 5.25). The new 3-step approach was also supported by the gain in usability from the reduced data linkage time (efficiency) and assists (effectiveness) and the lower values in the PSSUQ scales (satisfaction) compared to previous versions (Fig. 5.23). This statement was also supported by the codes related to the improved usability and output data in the PSSUQ comments (36 out of 120, Appendix E)

The usefulness of the output elements was emphasised by the experts. The data visualisation component was separated from the data linkage process in the form of an interactive report that could be explored online. The participants found the report helpful to explore and understand the data (169 out of 796, Fig. 5.25) with a clear presentation of the data as plots (112 out of 796, Fig. 5.25). Overall, the participants found the output complete to start their analysis with the provided (meta)data as CSV files (99 out of 796, Fig. 5.25).

Recommendations:

- P3.F1.1. Continue with the 3-step approach for the data linkage process.
- P3.F1.2. Maintain the data visualisation component as a UI and an exportable interactive report.

Theme: Potential uptake

Finding (P3.F2). *"SERDIF can be applied for real use cases with tailored environmental data and features even for non-technical researchers."*

Evidence. The participants mentioned that SERDIF would be useful for their particular or their colleagues' research (189 out of 1616, Table 5.9 and Fig. 5.25). The willingness to use the framework for research was expressed in different ways. The majority of the participants (17 out of 23, Fig. 5.25) suggested some additional features that would benefit their particular research or research group's approach. For example, selecting specific variables in the linkage process, including feedback on how to get more data coverage or adding sanity checks for the retrieved data as well as checking the health events input data for formatting mistakes (63 out of 189, Fig. 5.25, and 12 out of 120, Appendix E). Some of the participants were explicit about the tool and data applicability to real use cases (54 out of 189, Fig. 5.25) and the importance of data protection aspects built into the framework (20 out of 189, Fig. 5.25). In addition, the participants identified additional sources of environmental data, with different types, granularities and coverages, that they would like to link with health events.

The potential usefulness of the framework was also supported by the high satisfaction (i.e. low values close to 1) reflected in the PSSUQ scores and scales (Fig. 5.22) in meeting the expert requirements. The IQR range below 1 point (14 out of 19 questions, Fig. 5.22) indicating that the participants agreed on the overall usability as a group of experts. Even researchers that were not used to linking the actual data themselves, due to their responsibilities in the research group, noticed the potential usefulness of SERDIF for their colleagues (25 out of 189, Fig. 5.25).

Recommendations:

- P3.F2.1. Explore ways to manage the future development of the SERDIF implementation.
- P3.F2.2. Promote the framework and implementation with research groups studying environmental factors associated with health outcomes.
- P3.F2.3. Identify additional relevant data sources and import them to the KG.

Theme: Minor improvements

Finding (P3.F3). *"Some important features and text descriptions are not clear in the User Interface (UI)."*

Evidence. The second most referenced theme in the third usability test was Minor improvements (429 out of 1616, Table 5.9 and Fig. 5.25). The participants were not completely clear with some of the elements to upload or link the data at least once, as the most referenced code in this theme reflects (144 out of 429, Fig. 5.25). For example, participants uploaded the data twice by using the drag and drop function, and by clicking the upload button (Fig. 5.18); or partially selected some of the linkage options (Fig. 5.19). The participants also stated that some words and features needed clarification with tooltips or extra sentences in the text area (111 out of 429, Fig. 5.25, and 13 out of 120, Appendix E). Furthermore, the amount of information in the output reduced the understanding of the output content such as the data table and plot visualisations in the interactive report (148 out of 429, Fig. 5.25).

The references for the codes in this theme were distributed heterogeneously as denoted by the clusters in the heatmap highlighting few participants (P2, P18, P19 in Fig. 5.25). These three participants also took the longest while completing T1 (>20 min in Fig. 5.21). The improvements were considered minor due to the heterogeneous distributions and the lower total number of references compared to the *Requirements achieved* theme, with complementary codes to this theme.

Recommendations:

- P3.F3.1. Build in a check for the input health events data format.
- P3.F3.2. Simplify and clarify some of the text information in the UI and interactive report.

Theme: Testing methodology

Finding (P3.F4). *"The lack of preparation from the participants increased the need for the moderator's guidance."*

Evidence. The overall testing methodology has improved across the three phases denoted by the decrease in the total moderator assists (Table 5.9). The expert participants were able to complete the data linkage tasks (T1 and T2) with less than 6 assists in total (17 out of the 23 participants, Fig. 5.25). Even though there were less assists, more than half of them were due to the navigation and content being complex (63 out of 112, Fig. 5.24). However, the assists were considered as guidance rather than an assist in this Phase 3. That is because the assists were related to interventions mainly because the participant asks a direct question, wanders off task or the system had a technical issue.

Almost half of the participants had a technical issue while completing the tasks (Fig. 5.24), which was related to an invalid input at the upload step (Fig. 5.18) or an incompatible browser setting. The majority of the participants (16 out of 23, Fig. 5.25) was unsure if they had completed the task, had to read the tasks after some time and mentioned their lack of preparation for the session. Furthermore, most of the participants used the example events provided in the upload step of the UI (Fig. 5.18) without understanding what the events were related to nor the spatiotemporal context of the event. Despite the initial confusion, the experts managed to complete the tasks using SERDIF effectively as denoted by the improvement on the data linkage time and assists required (Fig. 5.23A, B).

Recommendations:

- P3.F4.1. Add a statement after each task providing the final action to understand if a task is finished.

- P3.F4.2. Add a pre-task where participants build their own health events dataset.

5.4.4.6 Conclusions

The third iteration of SERDIF evaluated with **Use Case 3 - AAV in Europe** yielded a satisfactory outcome, enabling researchers to link health events and environmental data using KG.

The summative usability part of the evaluation demonstrates evidence for the improvement of the usability of SERDIF against previous Phases. The efficiency, effectiveness and satisfaction aspects have improved in Phase 3 compared to Phase 2. Furthermore, the additional functionalities and larger sample size had a positive effect on SERDIF's measured usability.

The formative usability part of the evaluation (observational findings) support SERDIF in being usable for non-technical users and potentially useful for health-environment researchers (P3.F1 and P3.F2). However, the combination of some important data linkage features and text descriptions being unclear (P3.F1 and P3.F3) and the lack of preparation from the participants (P3.F4) required guidance from the moderator to complete the tasks.

The User-Centred Design (UCD) provided a more in depth understanding of the context of use which was translated into a successful refinement of the expert user requirements. The mixed methods approach, combining thematic analysis with quantitative metrics to support the findings and the transparency of the reporting of these findings, provided the effective means to refine the framework efficiently. The usability improvement throughout study reinforces the adequacy of the evaluation approach taken to evaluate the framework against the requirements.

5.5 Limitations

The design of the framework and tasks are focused on simplifying the formulation of SPARQL queries, exploration of the retrieved data (CSV) and the generation of interoperable data (RDF). Therefore, the design does not provide additional tools beyond the graph database (i.e. triplestore) functionalities to explore the KG, which limits the property and class exploration of the KG.

Another limitation of the design choice for SERDIF implementation is that environmental data needs to be manually collected by the researcher, and then uplifted to RDF with the help of a KG expert. The automatisation of the uplift process was not feasible since environmental data and metadata comes in different structures and formats, nor suitable endpoints with relevant environmental data for the use cases were available. A template R2RML mapping was provided to uplift the data in a semi-automatic manner to RDF following the QB structure, where the user can convert the environmental data to the template format, and then uplift the data. However, a KG expert is recommended to be part of the uplift process.

Regarding the subjectivity of our evaluation approach, the study followed best practices to minimise the subjectivity of the thematic analysis [Nowell et al., 2017]. The three authors participated in coding the data and reviewing the findings (Section 5.1.5), reducing the coding bias. The evidence for the findings combines multiple sources of qualitative and quantitative usability metrics (Section 5.1.4), lowering the findings bias. The participants were given the chance to review the results providing feedback to the authors, sustaining the representation of participants' views.

The usability study incorporated HDRs in a progressive manner from phase to phase. A total of 29 unique researchers participated in evaluating SERDIF with some of them present in more than one phase, which could lead to a potential training effect. The training effect could have introduced a bias for recurrent participants that may perform better in P2 and/or P3 due to their previous exposure to SERDIF in P1 and/or P2. The frequency of recurrent participants is 7 out of 29 for three phases, 7 out of

29 for two phases and 15 out of 29 for a single phase. Towards minimising the potential training effect, the usability study included HDRs from different research groups studying various diseases at different geographical levels, and increased the sample size throughout the phases. The recurrence of the participants in this usability study can indicate a positive reassurance of the potential usefulness of SERDIF as HDRs, often with busy schedules and little time for extra tasks, supported the evaluation of the framework.

The usability of SERDIF was only evaluated for domain experts, in particular Health Data Researchers (HDR), conducting data linkage tasks for the study of environmental factors associated with rare diseases. For the framework to be usable for domain experts in other domains or lay-users, further studies need to be conducted extending the type of participants in the usability evaluation.

5.6 Evaluation Conclusions

This chapter presented and discussed the findings of the usability evaluation study to evaluate the usability and potential usefulness of SERDIF to link health events and environmental data for research. The usability study comprised three iterative usability tests (Phase 1, Phase 2 and Phase 3) where the expert user requirements and the SERDIF components were refined based on the results of each phase. The summative metrics progression (Fig. 5.23) indicates a gain in usability in terms of efficiency, effectiveness and satisfaction for the group of researchers that participated in the study in the context of data linkage. The observational findings from the formative part of the study are summarised for each of the usability phases as:

Phase 1. *"SERDIF is a viable approach to facilitate the health and environmental data linkage process for researchers, but the user requirements, components of the framework and testing methodology need to be refined."*

Phase 2. *"SERDIF holds promise to be a useful and usable data linkage framework for health-environmental research, but the user requirements and components of the framework related to the data linkage process need to be refined, and the design of the tasks simplified."*

Phase 3. *"SERDIF is usable for non-technical researchers and potentially useful for health-environmental researchers. The implementation of the framework components achieved the expert user requirements in conducting data linkage tasks in a rare disease context."*

By way of summary, the need for refinement of requirements and of SERDIF components across the phases is shown in Table 5.10. The table describes if a refinement was needed (black background) or not (white background) in the next phase both for the requirements (R1, R2, and R3) and the components of the framework (as per each of the methodology Steps 1-6). The methodology steps include the refinement for the actual step (denoted by the letter M) and for the components (letter G for the Knowledge Graph and letter U for the User Interface) updated as a result of implementing the methodology step. In addition, the updates on the testing methodology are included as the letter T.

Table 5.10: Summary of the refinements on the expert user requirements and the implementation of the SERDIF components for each phase of the usability study. The black background indicates that the implementation of the component needs to be updated in the next phase, and the white background the opposite. The letters in the symbols denote the SERDIF components: M - Methodology, G - Knowledge Graph, U - User Interface and T - Testing methodology; and the requirements: 1 - Requirement 1, 2 - Requirement 2, and 3 - Requirement 3.

Usability test	**Requirements**	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**
Phase 1	❶❷❸	Ⓜ🅖	Ⓜ🅖	Ⓜ🅖	🅜🅤	🅜🅤	Ⓜ🅣
Phase 2	❶❷③	Ⓜ🅖	ⓂⒼ	Ⓜ🅖	Ⓜ🅤	🅜🅤	Ⓜ🅣
Phase 3	①②③	ⓂⒼ	ⓂⒼ	ⓂⒼ	Ⓜ🅤	ⓂⓊ	Ⓜ🅣

Chapter 6

Conclusions

This chapter draws conclusions from the research presented in this thesis. The extent to which the research objectives of this thesis, presented in Chapter 1, have been achieved is first discussed (Section 6.1). This chapter follows with the revisitation of the contributions of this thesis (Section 6.2). This chapter also discusses the impact and uptake from the research community (Section 6.3). This chapter follows with the potential future work topics based on the results of this thesis (Section 6.4). This chapter concludes with the final remarks of this thesis (Section 6.5).

6.1 Achievement of Research Objectives

The research question examined in this thesis, defined in Section 1.2, is:

To what extent can a Knowledge Graph (KG) framework, that is standards-based, enable Health Data Researchers (HDR) to effectively link environmental data with particular health events through location and time?

Four Research Objectives (RO) were defined in order to support the answering of the research question of this thesis (Section 1.3). This section discusses the extent of the achievement of the research objectives based on work presented in previous chapters of the thesis.

6.1.1 Achievement of Research Objective 1

RO1: Conduct a state-of-the-art review of how to make data interoperable and usable for scientific research.

This research objective was achieved through the analyses of the state of the art described in Chapter 3. The review first examined aspects of the challenge and suggested approaches in scientific literature on how to make data interoperable for research, towards an efficient data integration process. A collection of state-of-the-art *review papers* was analysed for general, health and other domains. Knowledge Graph (KG) approaches based on W3C standards emerged as the suggested solution from the analysis of the results to make data interoperable. However, researchers without practical experience in KGs (domain experts) have difficulties in including KGs in their research workflows due to a usability challenge. The review then explored how to facilitate the adoption of standards-based KG technologies for domain experts. Visualisation approaches are the preferred solution to enable the interaction with the interoperable data. In particular, a form-based query builder was highlighted from the review, emphasising the ease of use and ability to include and hide complex queries from non-technical users. While visual approaches have been used in the health domain, research on the usability aspect is limited to a few empirical usability evaluations that partially adopted best practices. Furthermore, none of the approaches represented the health-environmental or rare disease domains, providing an opportunity to make data interoperable and usable for researchers studying health outcomes associated with environmental factors.

The findings from the state of the art were explored in a preliminary study that evaluated a standards-based KG approach to facilitate the data integration task of Health Data Researchers (HDR) studying the environmental triggers of a particular rare disease. The results from the preliminary study motivated the design of a multi-component framework beyond a unique visualisation tool.

6.1.2 Achievement of Research Objective 2

RO2: Identify HDR requirements in linking data for health-environmental research.

This research objective was achieved during the development of the framework (SERDIF) following a User Centred Design (UCD) described in Chapter 3 and 5. An initial set of requirements was distilled after understanding the context of use referent to the AAV in Ireland use case defined in Section 5.2.1 and undertaking a consensus process with HDRs from the research project involved in this use case. The expert user requirements were refined in an iterative manner through the usability evaluation described in Chapter 5. The refinement included different research cultures from diverse groups and projects at a European level, generalising the requirements in each iteration. The resulting expert user requirements for HDRs trying to link health events with environmental data for research are as follows:

Requirement 1 (R1.2). *Enable HDRs to query environmental data associated with relevant/own individual health events through location and time, within the area of the event and a period of data before the event.*
Requirement 2 (R2.2). *Support the understanding of event-environmental linked data and metadata, with its use, limitations and data protection risk for individuals, by using a simplified view focused on the data linkage process with optional further information.*
Requirement 3 (R3.2). *Export event-environmental linked (meta)data to be used as input in statistical models for data analysis (CSV) and for publication (CSV, RDF).*

6.1.3 Achievement of Research Objective 3

RO3: Develop a framework that enables a HDR to link environmental data with particular health events based on user data inputs.

This research objective was achieved by developing the Semantic Environmental and Rare Disease Integration Framework (SERDIF) described in Chapter 4. SERDIF was informed by the state-of-the-art findings and designed to achieve the user expert requirements from three use cases in the health-environmental domain. The framework is a combination of three components: a methodology, KG and a User Interface (UI).

Based on state-of-the-art findings, SERDIF includes a methodology to develop a usable KG approach based on W3C standards for an effective data integration process in the health-environmental domain. The methodology component also included a usability evaluation step based on best practices to validate the implementation against a particular use case following a UCD. Based on expert user requirements, SERDIF was designed to enable researchers in performing data linkage tasks that relied on health events and linkage parameters relevant for their studies. The framework combined a

W3C standard KG implementation to address the data interoperability challenge and a UI to interact with the KG in a usable manner.

6.1.4 Achievement of Research Objective 4

RO4: Evaluate and refine the developed framework through rare disease case studies.

The expert user requirements to design and develop a framework to link health and environmental data were presented in Chapter 4. The achievement of the expert user requirements by SERDIF is summarised below.

Requirement 1 (R1.2). *Enable HDRs to query environmental data associated with relevant/own individual health events through location and time, within the area of the event and a period of data before the event.*

The design of a framework to link health events and environmental data achieved this requirement. SERDIF includes a UI component that allows HDRs to build complex queries in an intelligible manner. Researchers can link health events and environmental data through location and time, which are common dimensions in both data types. The UI also hides the graph data model used to structure the environmental data in the KG.

Requirement 2 (R2.2). *Support the understanding of event-environmental linked data and metadata, with its use, limitations and data protection risk for individuals, by using a simplified view focused on the data linkage process with optional further information.*

The combination of a simplified UI to perform the data linkage process, structured in three separate sections, and an interactive report exported from the UI achieved this requirement. The sections in the UI included (i) health event input or upload (R1), (ii) data linkage options (R1) and (iii) export output (R2, R3). The interactive report as an HTML file made the linked data and metadata easier to understand before starting the analysis, including a section to explore and visualise the data within environmental context.

Requirement 3 (R3.2). *Export event-environmental linked (meta)data to be used as input in statistical models for data analysis (CSV) and for publication (CSV, RDF).*

The export output from the SERDIF UI included the linked event-environmental data for analysis as a data table (CSV) and graph (RDF), the metadata describing the linkage process and the data (CSV and RDF) and the interactive report mentioned above. The CSV data was structured and content was acknowledged to be usable as input for HDRs workflows. The RDF data is ready to be published in an open data repository if it is not considered as personal data.

SERDIF met the expert user requirements gathered from undertaking a consensus process with Health Data Researchers (HDRs), supported by the findings from the state-of-the-art reviews in data interoperability and KG usability challenges in health-environmental studies, and refined in an iterative process as part of a User-Centred Design (UCD).

6.2 Contributions

This section briefly revisits the contributions from the research of this thesis, which were initially presented in Chapter 1. The research of this thesis resulted in three contributions: one major and two minor contributions.

The **major contribution** of this thesis is the **Semantic Environmental and Rare Disease data Integration Framework (SERDIF)**, comprising Knowledge Graph (KG), Methodology and User Interface (UI). The framework enables Health Data Researchers (HDR) to link health events with relevant environmental data through location and time. The evaluation results indicate that SERDIF is usable and potentially useful for HDRs conducting data integration tasks in the health-environmental domain.

SERDIF advances the state of the art in being the first usable standards-based KG approach to be developed and implemented for the study of environmental triggers associated with rare diseases. The User Centred Design (UCD) provided the framework with the possibility to incorporate expert feedback throughout the development process, which promoted the collaboration between domain experts and KG practitioners towards achieving a shared goal. SERDIF also promoted the transparency of the linkage process to facilitate the understanding of the linked health-environmental data for researchers.

The resulting linked data is provided with enough information about the origin of the data and processing steps in a human- (CSV and HTML) and machine-understandable (RDF) format, following best practices for data on the Web from W3C.

The implementation of the framework results in linked data ready to be used in the researcher's workflows and published as Open Data to be reused by other researchers in different contexts. Furthermore, the framework is developed to comply with and promote the data governance aspect of the processing of health and environmental data, central to the linkage process. The linked data is ready to be deposited in an open data repository towards making the data Findable, Accessible, Interoperable and Reusable (FAIR). The achievement of FAIR data practices and goals can benefit future European and International projects with data linkage tasks present in their agendas. These projects can follow the technology independent design of the methodology component of SERDIF or adopt the W3C standard approach design choice based on their goals and context.

SERDIF has the potential to be used in other contexts and domains to address the data integration challenges of environmental studies providing the bases to inform professionals in decision making.

The only requirement from the input health event data is to have a date and location to be linked with environmental data. Therefore, dates can also represent populations such as a peak in flu cases in a specific city, county or country. Expanding the application possibilities of the framework to study environmental factors linked to any disease or health event.

Furthermore, (i) ecological, (ii) sociological, (iii) political, (iv) sustainable business environmental and (v) pharmacological studies could benefit from our research. For example, (i) when studying the environmental conditions a population of animals or plants have been exposed to; (ii) comparing survey results on the perceived quality of the environment with actual environmental data; (iii) advocating with evidence that certain communities are exposed to poor environmental conditions; (iv) providing a record of the air quality in the business surroundings to demonstrate their impact on the environment; (v) gathering complementary from the environment to inform the risk assessment of a particular drug towards saving money before clinical trials.

Besides the direct impact on environmental studies, SERDIF has also the potential to become the foundation of an early warning system for public health researchers investigating outbreaks, including potential future global pandemics, such as COVID-19.

The **first minor contribution** of this thesis is **a step by step description of the methods and results of the evaluation approach**. The evaluation has been proven to be effective in improving the usability and potential usefulness of each of the components of SERDIF (i.e. KG, Methodology and UI). The methods and results of the evaluation advance the state of the art in being the first usability results published for a standards-based KG approach in the health-environmental domain, following best practices in conducting usability studies. Beyond the contribution to the health-environmental domain, the detailed description and implementation of the evaluation approach can provide KG practitioners with reproducible research to initiate themselves in conducting usability studies towards improving the adoption of KG technologies. For example, research and industry projects might want to benefit from representing their data as a KG, to facilitate the integration and contextualisation of their data, but the data will be explored and analysed by experts without a background in KG technologies. Therefore, making the data usable for experts could be time and cost effective for their projects.

The **second minor contribution** of this thesis is the collection of **open source artefacts** as a by-product during the development of the framework. This is a contribution to open science promoting transparency of research methods and data reuse towards improving the efficiency of scientific research. The open source artefacts serve as support for the other two contributions of this thesis. The open data datasets published exemplify the KG input and output data, which are the uplifted environmental and geometry data as RDF and the linked health-environmental data resulting from a SERDIF query. The code published in this thesis' GitHub repository grants the possibility to other practitioners to fork the framework and adapt it to their own use case, while being able to reproduce the usability results and reusing the data analysis scripts.

6.3 Impact and uptake

This section presents the impact from the research of this thesis in the scientific community.

This research is already showing the impact within the research community through successful publications arising out of the research in top Web venues such as VOILA, the premier venue for HCI issues related to KGs; International Joint Conference on Knowledge Graphs (IJCKG), a premium academic and international forum for KGs; and Semantic Web journal, a top journal for interoperability, usability and applicability of Semantic Web technologies. Other general and health related venues where this research has been published include the Marie Curie Annual Conference (MCAA) and International Vasculitis and ANCA Workshop, the largest international congress focused on vasculitis disorders.

Researchers involved in the three use cases presented in Chapter 5 from the AVERT, HELICAL and FAIRVASC projects and from the Climate&Health group showed interest in incorporating SERDIF as part of their health-environmental workflows.

Furthermore, the author from this thesis participated in public engagement activities to increase the outreach for secondary school students, postgraduates in health sciences, patient cohorts and employees of a company. The following list includes the activities conducted during the development of this thesis:

- Four pub quizzes to raise the awareness of rare diseases in collaboration with Vasculitis Ireland Awareness[1] (patient cohort), Northern Ireland rare disease partnership[2] (high school students and postgraduates in health sciences) and Vifor[3] (company employees).
- Researchers' Night 2021: an interactive game was designed to present how researchers apply data protection regulations to work with rare disease data for research as the patients' identities need to remain a secret[4].
- Researchers' Night 2022: an interactive game to demonstrate how medical professionals work with computer scientists to better understand how the quality of the air around us affects our health[5].

6.4 Future Work

This section discusses the potential future work that could be undertaken based on the findings of this thesis.

Regarding the usability of the framework, further research could develop an extension to the framework with the focus on patient cohorts. The tasks for the patients could include annotating the KG with relevant health related information requested by a health professional. For example, patients could include diet, exercise, sleep or stool deposition information in a self-reported manner. This information could be combined with existing health and environmental data for a more complete view of the health event, towards following the effects of treatment on a patient. Furthermore, patients can be provided with the means to explore a summarised view of their weekly or monthly health parameters as an interactive visualisation. This approach can promote the involvement of patients with health experts and research projects, providing another view to the study and strengthen the community.

Regarding the evaluation of the framework, future research can consider testing the scalability, technical validity and data quality. Researchers can test the use of federated queries to integrate data from different endpoints to integrate the relevant datasets, even beyond environmental data. The datasets were requested to be integrated and transformed using particular arithmetic functions but due to the flexibility of the approach, practitioners can include other types of query patterns for different

[1] https://vasculitis-ia.org/

[2] https://nirdp.org.uk/

[3] https://www.viforpharma.ch/

[4] https://www.tcd.ie/research/start/dr-data.php

[5] https://www.adaptcentre.ie/news-and-events/adapt-programme-at-start-european-researchers-night-2022

use case requirements. The KG approach can be tested against other data integration approaches to find a more optimal solution, which could combine KG technologies with other approaches. A data quality assessment could be generated during the uplift process and after the linkage query to complement the framework with a test for the fitness of the data to be used given a particular context of use.

The use cases evaluated in this thesis included ongoing rare disease research that are at early stages due to the complexity of the disease and availability of the data. However, future studies on use cases that are more advanced can include logic rules to look for specific patterns across the data. The patterns can be also co-designed with domain experts towards developing an expert system capable of discovering new links in an automatic manner. The system could search the KG for certain combinations of parameters, by using SPARQL query templates, that can inform the risk of people having a health event, improving the quality of life for people with certain conditions.

Furthermore, the expert system has the potential to set the grounds for an early warning system to be used EU wide. The expert system would need to include an automatic data uplift stream to import the up-to-date and forecast data from the environmental sites and a connection with the healthcare centres to assess the health risk for specific vulnerable groups; and obtain the necessary approvals to be used as a risk assessment system in terms of GDPR and local regulations. The system will provide the necessary documentation for healthcare centres and individuals to trace how the risk warning was informed. However, the dissemination and communication channels to inform the individuals are envisioned to be managed by the healthcare centres or local authorities, as well as, the preparation and formation of the individuals on how to respond to a risk warning.

6.5 Final Remarks

It is hoped by the author of this thesis that SERDIF, a framework to effectively health events with environmental data, can be of benefit to researchers studying the environmental factors associated with particular health events; and to researchers looking to make emerging technologies more usable.

It is also hoped by the author of this thesis that SERDIF would benefit the research community. Researchers can incorporate SERDIF to facilitate the data integration tasks in their workflows, use the findings in this thesis in their research, and apply their expertise to contribute to the approach and its implementations.

References

Abedjan, Z., Boujemaa, N., Campbell, S., Casla, P., Chatterjea, S., Consoli, S., Costa-Soria, C., Czech, P., Despenic, M., Garattini, C., Hamelinck, D., Heinrich, A., Kraaij, W., Kustra, J., Lojo, A., Sanchez, M. M., Mayer, M. A., Melideo, M., Menasalvas, E., Aarestrup, F. M., Artigot, E. N., Petković, M., Recupero, D. R., Gonzalez, A. R., Kerremans, G. R., Roller, R., Romao, M., Ruping, S., Sasaki, F., Spek, W., Stojanovic, N., Thoms, J., Vasiljevs, A., Verachtert, W., and Wuyts, R. (2019). Data science in healthcare: Benefits, challenges and opportunities. In Consoli, S., Reforgiato Recupero, D., and Petković, M., editors, *Data Science for Healthcare: Methodologies and Applications*, pages 3–38. Springer International Publishing. https://doi.org/10.1007/978-3-030-05249-2_1.

Afsar, B., Elsurer Afsar, R., Kanbay, A., Covic, A., Ortiz, A., and Kanbay, M. (2019). Air pollution and kidney disease: review of current evidence. 12(1):19–32. https://doi.org/10.1093/ckj/sfy111.

Aguiar, M., Nunez, S., and Giesteira, B. (2021). A survey on user interaction with linked data. In Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Sixth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 3023 of *VOILA'21*, pages 13–28, Virtual Workshop. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3023/paper5.pdf.

Al-Kindi, S. G., Brook, R. D., Biswal, S., and Rajagopalan, S. (2020). Environmental determinants of cardiovascular disease: lessons learned from air pollution. 17(10):656–672. https://doi.org/10.1038/s41569-020-0371-2.

Al-Tawil, M., Dimitrova, V., and Thakker, D. (2020). Using knowledge anchors to facilitate user exploration of data graphs. 11(2):205–234. https://doi.org/10.3233/SW-190347.

Albertoni, R., Browning, D., Cox, S. J. D., Gonzalez-Beltran, A., Perego, A., and Winstanley, P. (2020). Data Catalog Vocabulary (DCAT) - Version 2. W3c recommendation, W3C. https://www.w3.org/TR/vocab-dcat-2/.

Ali, S., Zada, I., Mehmood, Z., Ullah, A., Ali, H., and Ullah, M. (2022). Publishing and interlinking COVID-19 data using linked open data principles: Toward effective healthcare planning and decision-making. 2022:e4792909. https://doi.org/10.1155/2022/4792909.

Allen, R. and Hartland, D. (2018). FAIR in practice - jisc report on the findable accessible interoperable and reuseable data principles. Technical report, JISC. `https://doi.org/10.5281/zenodo.1245568`.

Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C., and Yang, G.-Z. (2015). Big data for health. 19(4):1193–1208. `https://doi.org/10.1109/JBHI.2015.2450362`.

Angles, R., Arenas, M., Barceló, P., Hogan, A., Reutter, J., and Vrgoč, D. (2017). Foundations of modern query languages for graph databases. 50(5):68:1–68:40. `https://doi.org/10.1145/3104031`.

Arenas, M., Bertails, A., Prud'hommeaux, E., and Sequeda, J. (2012). A direct mapping of relational data to RDF. W3c recommendation, W3C. `https://www.w3.org/TR/rdb-direct-mapping/`.

Atemezing, G., Corcho, O., Garijo, D., Mora, J., Poveda-Villalón, M., Rozas, P., Vila-Suero, D., and Villazón-Terrazas, B. (2013). Transforming meteorological data into linked data. 4(3):285–290. `https://doi.org/10.3233/SW-120089`.

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., Bernal-Delgado, E., Blomberg, N., Bock, C., Conesa, A., Del Signore, S., Delogne, C., Devilee, P., Di Meglio, A., Eijkemans, M., Flicek, P., Graf, N., Grimm, V., Guchelaar, H.-J., Guo, Y.-K., Gut, I. G., Hanbury, A., Hanif, S., Hilgers, R.-D., Honrado, , Hose, D. R., Houwing-Duistermaat, J., Hubbard, T., Janacek, S. H., Karanikas, H., Kievits, T., Kohler, M., Kremer, A., Lanfear, J., Lengauer, T., Maes, E., Meert, T., Müller, W., Nickel, D., Oledzki, P., Pedersen, B., Petkovic, M., Pliakos, K., Rattray, M., i Màs, J. R., Schneider, R., Sengstag, T., Serra-Picamal, X., Spek, W., Vaas, L. A. I., van Batenburg, O., Vandelaer, M., Varnai, P., Villoslada, P., Vizcaíno, J. A., Wubbe, J. P. M., and Zanetti, G. (2016). Making sense of big data in health research: Towards an EU action plan. 8(1):71. `https://doi.org/10.1186/s13073-016-0323-y`.

Augustin, H., Sudmanns, M., Tiede, D., Lang, S., and Baraldi, A. (2019). Semantic earth observation data cubes. 4(3):102. `https://doi.org/10.3390/data4030102`.

AVERT, p. (2022). Rare kidney disease registry and biobank. `https://www.tcd.ie/medicine/thkc/avert/`.

Bartolomeo, S. D., Pepe, G., Savo, D. F., and Santarelli, V. (2018). Sparqling: Painlessly drawing SPARQL queries over graphol ontologies. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Fourth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 2187 of *CEUR Workshop Proceedings*, pages 64–69. CEUR. `https://ceur-ws.org/Vol-2187/#paper6`.

Belete, G. F., Voinov, A., and Laniak, G. F. (2017). An overview of the model integration process: From pre-integration assessment to testing. 87:49–63. https://doi.org/10.1016/j.envsoft.2016.10.013.

Bikakis, N. and Sellis, T. K. (2016). Exploration and visualization in the web of big linked data: A survey of the state of the art. *ArXiv*.

Biron, P. V. and Malhotra, A. (2004). Xml schema part 2: Datatypes second edition. W3c recommendation, W3C. https://www.w3.org/TR/xmlschema-2/.

Blinkiewicz, M. and Bak, J. (2016). SQuaRE: A visual support for OBDA approach. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Second International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 1704 of *CEUR Workshop Proceedings*, pages 41–53. CEUR. https://ceur-ws.org/Vol-1704/#paper4.

Boren, T. and Ramey, J. (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3):261–278. https://doi.org/10.1109/47.867942.

Bottoni, P. and Ceriani, M. (2015). SPARQL playground: A block programming tool to experiment with SPARQL. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the International Workshop on Visualizations and User Interfaces for Ontologies and Linked Data*, volume 1456 of *CEUR Workshop Proceedings*, page 103. CEUR. https://ceur-ws.org/Vol-1456/#paper12.

Boyles, R. R., Thessen, A. E., Waldrop, A., and Haendel, M. A. (2019). Ontology-based data integration for advancing toxicological knowledge. 16:67–74. https://doi.org/10.1016/j.cotox.2019.05.005.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. 3(2):77–101. https://doi.org/10.1191/1478088706qp063oa.

Braşoveanu, A. M. P., Sabou, M., Scharl, A., Hubmann-Haidvogel, A., and Fischl, D. (2017). Visualizing statistical linked knowledge for decision support. 8(1):113–137. https://doi.org/10.3233/SW-160225.

Brickley, D. and Guha, R. (2014). Rdf schema 1.1. W3c recommendation. https://www.w3.org/TR/rdf-schema/.

Brooke, J. (1996). SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*, 189:189–194.

Bruyat, J., Champin, P.-A., Médini, L., and Laforest, F. (2022). Shacled turtle: Schema-based turtle auto-completion. In Fu, B., Lambrix, P., and Pesquita, C., editors, *Proceedings of the Seventh International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3253 of *CEUR Workshop Proceedings*, pages 2–15. CEUR. https://ceur-ws.org/Vol-3253/#paper1.

Bălă, G.-P., Râjnoveanu, R.-M., Tudorache, E., Motișan, R., and Oancea, C. (2021). Air pollution exposure—the (in)visible risk factor for respiratory diseases. 28(16):19615–19628. https://doi.org/10.1007/s11356-021-13208-x.

Callahan, T. J., Tripodi, I. J., Pielke-Lombardo, H., and Hunter, L. E. (2020). Knowledge-based biomedical data science. 3:23–41. https://doi.org/10.1146/annurev-biodatasci-010820-091627.

Cambridge, D. (2023). research agenda collocation | meanings and examples of use.

Canali, S. and Leonelli, S. (2022). Reframing the environment in data-intensive health sciences. 93:203–214. https://doi.org/10.1016/j.shpsa.2022.04.006.

Chakraborty, J., Padki, A., and Bansal, S. K. (2017). Semantic ETL — state-of-the-art and open research challenges. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, ICSC'17, pages 413–418. IEEE. https://doi.org/10.1109/ICSC.2017.94.

Chaves-Fraga, D. and Dimou, A. (2022). Declarative description of knowledge graphs construction automation: Status & challenges david chaves-fraga, anastasia dimou. In Chaves-Fraga, D., Dimou, A., Heyvaert, P., Priyatna, F., and Sequeda, J., editors, *Proceedings of the 3rd International Workshop on Knowledge Graph Construction (KGCW 2022)*, volume 3141 of *CEUR Workshop Proceedings*. CEUR. https://ceur-ws.org/Vol-3141/paper5.pdf.

Choquet, R., Maaroufi, M., Fonjallaz, Y., de Carrara, A., Vandenbussche, P.-Y., Dhombres, F., and Landais, P. (2015). LORD: a phenotype-genotype semantically integrated biomedical data tool to support rare disease diagnosis coding in health information systems. 2015:434–440. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765596/.

Christofidou, M., Lea, N., and Kalra, D. (2023). Helical data protection impact assessment (dpia) template. Technical report. https://cordis.europa.eu/project/id/813545/results.

Coccia, M. and Benati, I. (2018). Comparative studies. In Farazmand, A., editor, *Global Encyclopedia of Public Administration, Public Policy, and Governance*, pages 1–7. Springer International Publishing. https://doi.org/10.1007/978-3-319-31816-5_1197-1.

Commission, E. (2022). Horizon europe (HORIZON) programme guide - section open science. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf.

Copernicus, C. C. S. (2020). E-OBS daily gridded meteorological data for europe from 1950 to present derived from in-situ observations. https://doi.org/10.24381/CDS.151D3EC6.

Courtot, M. (2021). Data use ontology (DUO). Technical report. https://github.com/EBISPOT/DUO.

Cox, S. and Little, C. (2022). Time ontology in owl. W3c candidate recommendation, W3C. https://www.w3.org/TR/owl-time/.

Crotti Junior, A., Debruyne, C., Brennan, R., and O'Sullivan, D. (2017a). An evaluation of uplift mapping languages. 13(4):405–424. https://doi.org/10.1108/IJWIS-04-2017-0036.

Crotti Junior, A., Debruyne, C., and O'Sullivan, D. (2017b). Juma: An editor that uses a block metaphor to facilitate the creation and editing of r2rml mappings. In Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., and Hartig, O., editors, *The Semantic Web: ESWC 2017 Satellite Events*, Lecture Notes in Computer Science, pages 87–92. Springer International Publishing. https://doi.org/10.1007/978-3-319-70407-4_17.

Cui, Y., Balshaw, D. M., Kwok, R. K., Thompson, C. L., Collman, G. W., and Birnbaum, L. S. (2016). The exposome: Embracing the complexity for discovery in environmental health. 124(8):A137–A140. https://doi.org/10.1289/EHP412.

Cyganiak, R. and Reynolds, D. (2014). The RDF Data Cube Vocabulary. W3c recommendation, W3C. https://www.w3.org/TR/vocab-data-cube/.

Cyganiak, R., Wood, D., and Lanthaler, M. (2014). Rdf 1.1 concepts and abstract syntax. W3c recommendation. https://www.w3.org/TR/rdf11-concepts/.

Dadzie, A.-S. and Pietriga, E. (2016). Visualisation of linked data – reprise. 8(1):1–21. https://doi.org/10.3233/SW-160249.

Dadzie, A.-S. and Rowe, M. (2011). Approaches to visualising linked data: A survey. 2(2):89–124. https://doi.org/10.3233/SW-2011-0037.

Dafli, E., Antoniou, P., Ioannidis, L., Dombros, N., Topps, D., and Bamidis, P. D. (2015). Virtual Patients on the Semantic Web: A Proof-of-Application Study. *Journal of Medical Internet Research*, 17(1):e16. https://www.jmir.org/2015/1/e16/.

Das, S., Sundara, S., and Atkinson, R. (2012). R2rml: RDB to RDF mapping language. W3c recommendation, W3C. https://www.w3.org/TR/r2rml/.

DCMI, U. B. (2020). Dcmi metadata terms (dct). https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

De Meester, B., Heyvaert, P., and Delva, T. (2022). RDF mapping language (RML). Future w3c recommendation, rml.io. https://rml.io/specs/rml/.

de Mello, B. H., Rigo, S. J., da Costa, C. A., da Rosa Righi, R., Donida, B., Bez, M. R., and Schunke, L. C. (2022). Semantic interoperability in health records standards: a systematic literature review. 12(2):255–272. https://doi.org/10.1007/s12553-022-00639-w.

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., and Pentland, A. 2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. 347(6221):536–539. https://doi.org/10.1126/science.1256297.

De Santo, A. and Holzer, A. (2020). Interacting with linked data: A survey from the SIGCHI perspective. In *Extended Abstracts of the 2020 CHI Conference on*

Human Factors in Computing Systems, CHI EA '20, pages 1–12. Association for Computing Machinery. https://doi.org/10.1145/3334480.3382909.

Debattista, J., Auer, S., and Lange, C. (2016). Luzzu – a framework for linked data quality assessment. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 124–131. https://doi.org/10.1109/ICSC.2016.48.

Debruyne, C. and O'Sullivan, D. (2016). R2rml-f: Towards sharing and executing domain logic in r2rml mappings. In *Proceedings of the International Workshop on Linked Data on the Web, LDOW2016, co-located with the 25th International World Wide Web Conference (WWW 2016)*. https://ceur-ws.org/Vol-1593/article-13.pdf.

Debruyne, C., Walshe, B., and O'Sullivan, D. (2015). Towards a project centric metadata model and lifecycle for ontology mapping governance. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, iiWAS '15, pages 1–10. Association for Computing Machinery. https://doi.org/10.1145/2837185.2837201.

Desimoni, F., Bikakis, N., Po, L., and Papastefanatos, G. (2020). Linked data visualization tools. In *Linked Data Visualization*, pages 47–72. Springer International Publishing. https://doi.org/10.1007/978-3-031-79490-2_3.

Desolda, G., Matera, M., and Lanzilotti, R. (2020). Metamorphic data sources: A user-centric paradigm to consume linked data in interactive workspaces. 102:992–1015. https://doi.org/10.1016/j.future.2019.09.032.

Dhayne, H., Haque, R., Kilany, R., and Taher, Y. (2019). In search of big medical data integration solutions - a comprehensive survey. 7:91265–91290. https://doi.org/10.1109/ACCESS.2019.2927491.

Dimou, A., Kontokostas, D., Freudenberg, M., Verborgh, R., Lehmann, J., Mannens, E., Hellmann, S., and Van de Walle, R. (2015). Assessing and refining mappingsto RDF to improve dataset quality. In Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., and Staab, S., editors, *The Semantic Web - ISWC 2015*, Lecture Notes in Computer Science, pages 133–149. Springer International Publishing. https://doi.org/10.1007/978-3-319-25010-6_8.

Dimou, A., Nies, T. D., and Verborgh, R. (2016). Automated metadata generation for linked data generation and publishing workflows. In *Proceedings of the International Workshop on Linked Data on the Web, LDOW2016, co-located with the 25th International World Wide Web Conference (WWW 2016)*. https://ceur-ws.org/Vol-1593/article-04.pdf.

Drury, B., Fernandes, R., Moura, M.-F., and de Andrade Lopes, A. (2019). A survey of semantic web technology for agriculture. 6(4):487–501. https://doi.org/10.1016/j.inpa.2019.02.001.

EEA (2022). Eea air quality data. https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm/.

Ehrhart, T., Lisena, P., and Troncy, R. (2021). KG explorer: a customisable exploration tool for knowledge graphs. In Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Sixth International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3023 of *CEUR Workshop Proceedings*, pages 63–75. CEUR. https://ceur-ws.org/Vol-3023/#paper13.

EPA (2023). Airquality.ie. https://airquality.ie/.

EPA, O. (2015). Environmental protection agency (EPA), OMS: Environmental data. https://www.epa.gov/quality/about-managing-quality-environmental-data-epa-region-3.

EU (2016). General data protection regulation (gdpr) - regulation (eu) 2016/679 of the european parliament and of the council. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

EUPATI (2023). European patients' academy on therapeutic innovation(EUPATI): health event. https://toolbox.eupati.eu/glossary/health-event/.

eurostat (2021). Background - NUTS - nomenclature of territorial units for statistics - eurostat. https://ec.europa.eu/eurostat/web/nuts/background.

Euzenat, J. and Shvaiko, P. (2013). *Ontology Matching*. Springer Berlin, Heidelberg, 2 edition. https://doi.org/10.1007/978-3-642-38721-0.

Evans, K. J., Terhorst, A., and Kang, B. H. (2017). From data to decisions: Helping crop producers build their actionable knowledge. 36(2):71–88. https://doi.org/10.1080/07352689.2017.1336047.

FAIRVASC, p. (2022). Vasculitis registries across europe. https://fairvasc.eu/.

Farias Lóscio, B., Burle, C., and Calegari, N. (2017). Data on the web best practices. https://www.w3.org/TR/dwbp/.

Florenzano, F., Parra, D., Reutter, J. L., and Venegas, F. (2016). A visual aide for understanding endpoint data. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Second International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 1704 of *CEUR Workshop Proceedings*, pages 102–113. CEUR. https://ceur-ws.org/Vol-1704/#paper9.

Fuenmayor, L., Collarana, D., Lohmman, S., and Auer, S. (2017). FaRBIE: A faceted reactive browsing interface for multi RDF knowledge graph exploration. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Third International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 1947 of *CEUR Workshop Proceedings*, pages 111–122. CEUR. https://ceur-ws.org/Vol-1947/#paper10.

Gainotti, S., Torreri, P., Wang, C. M., Reihs, R., Mueller, H., Heslop, E., Roos, M., Badowska, D. M., de Paulis, F., Kodra, Y., Carta, C., Martìn, E. L., Miller, V. R., Filocamo, M., Mora, M., Thompson, M., Rubinstein, Y., Posada de la Paz, M., Monaco, L., Lochmüller, H., and Taruscio, D. (2018). The RD-connect registry & biobank finder: a tool for sharing aggregated data and metadata among rare disease researchers. 26(5):631–643. `https://doi.org/10.1038/s41431-017-0085-z`.

Gervasi, C., Ferrari, A., Papoutsis, I., and Touloumtzi, S. (2021). Deepcube: Explainable ai pipelines for big copernicus data. Technical report, GEOmedia. `https://mediageo.it/ojs/index.php/GEOmedia/article/view/1802`.

Gil, Y., Pierce, S. A., Babaie, H., Banerjee, A., Borne, K., Bust, G., Cheatham, M., Ebert-Uphoff, I., Gomes, C., Hill, M., Horel, J., Hsu, L., Kinter, J., Knoblock, C., Krum, D., Kumar, V., Lermusiaux, P., Liu, Y., North, C., Pankratius, V., Peters, S., Plale, B., Pope, A., Ravela, S., Restrepo, J., Ridley, A., Samet, H., Shekhar, S., Skinner, K., Smyth, P., Tikoff, B., Yarmey, L., and Zhang, J. (2018). Intelligent systems for geosciences: an essential research agenda. 62(1):76–84. `https://doi.org/10.1145/3192335`.

Giuliani, G., Masó, J., Mazzetti, P., Nativi, S., and Zabala, A. (2019). Paving the way to increased interoperability of earth observations data cubes. 4(3):113. `https://doi.org/10.3390/data4030113`.

Gligorijević, V. and Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. 12(112):20150571. `https://doi.org/10.1098/rsif.2015.0571`.

Golriz Khatami, S., Robinson, C., Birkenbihl, C., Domingo-Fernández, D., Hoyt, C. T., and Hofmann-Apitius, M. (2020). Challenges of integrative disease modeling in alzheimer's disease. 6. `https://doi.org/10.3389/fmolb.2019.00158`.

Grant, M. J. and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. 26(2):91–108. `https://doi.org/10.1111/j.1471-1842.2009.00848.x`.

Grapov, D., Fahrmann, J., Wanichthanarak, K., and Khoomrung, S. (2018). Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. 22(10):630–636. `https://doi.org/10.1089/omi.2018.0097`.

Graux, D., Orlandi, F., Kaushik, T., Kavanagh, D., Jiang, H., Bredican, B., Grouse, M., and Geary, D. (2021). Timelining knowledge graphs in the browser (short paper). In Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Sixth International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3023 of *CEUR Workshop Proceedings*, pages 76–81. CEUR. `https://ceur-ws.org/Vol-3023/#paper11`.

Graux, D., Orlandi, F., Lynch, B., Mahon, I., Mullen, O., Mahon, A., Molnar, F., and Mantiquilla, L. (2020). A real-time visual dashboard for wikidata edits. In

Ivanova, V., Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 2778 of *CEUR Workshop Proceedings*, pages 41–46. CEUR. `https://ceur-ws.org/Vol-2778/#paper4`.

Group, W. O. W. (2012). Owl 2 web ontology language document overview (second edition). W3c recommendation. `https://www.w3.org/TR/owl2-overview/`.

Haendel, M., Vasilevsky, N., Unni, D., Bologa, C., Harris, N., Rehm, H., Hamosh, A., Baynam, G., Groza, T., McMurry, J., Dawkins, H., Rath, A., Thaxon, C., Bocci, G., Joachimiak, M. P., Köhler, S., Robinson, P. N., Mungall, C., and Oprea, T. I. (2020). How many rare diseases are there? 19(2):77–78. `https://doi.org/10.1038/d41573-019-00180-y`.

Haller, A., Janowicz, K., Cox, S. J. D., Le Phuoc, D., Taylor, K., and Lefrançois, M. (2017). Semantic sensor network ontology. W3c recommendation, W3C. `https://www.w3.org/TR/vocab-ssn/`.

Haller, K., Ekaputra, F. J., Sabou, M., and Piroi, F. (2022). Enabling exploratory search on manufacturing knowledge graphs. In Fu, B., Lambrix, P., and Pesquita, C., editors, *Proceedings of the Seventh International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3253 of *CEUR Workshop Proceedings*, pages 16–28. CEUR. `https://ceur-ws.org/Vol-3253/#paper2`.

Hanlon, R., Barry, M., Marrinan, F., and O'Sullivan, D. (2021). Towards an effective user interface for data exploration, data quality assessment and data integration. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 431–436. `https://doi.org/10.1109/ICSC50631.2021.00077`.

He, X., Zhang, H., and Bian, J. (2020). User-centered design of a web-based crowdsourcing-integrated semantic text annotation tool for building a mental health knowledge base. *Journal of Biomedical Informatics*, 110:103571. `https://doi.org/10.1016/j.jbi.2020.103571`.

He, X., Zhang, R., Rizvi, R., Vasilakes, J., Yang, X., Guo, Y., He, Z., Prosperi, M., Huo, J., Alpert, J., and Bian, J. (2019). ALOHA: developing an interactive graph-based visualization for dietary supplement knowledge graph through user-centered design. *BMC Medical Informatics and Decision Making*, 19(4):150. `https://doi.org/10.1186/s12911-019-0857-1`.

Heacock, M. L., Amolegbe, S. M., Skalla, L. A., Trottier, B. A., Carlin, D. J., Henry, H. F., Lopez, A. R., Duncan, C. G., Lawler, C. P., Balshaw, D. M., and Suk, W. A. (2020). Sharing SRP data to reduce environmentally associated disease and promote transdisciplinary research. 35(2):111–122. `https://doi.org/10.1515/reveh-2019-0089`.

HELICAL, p. (2023). Health data linkage for clinical benefit. `https://doi.org/`

10.3030/813545.

Heyvaert, P., De Meester, B., Dimou, A., and Verborgh, R. (2019). Rule-driven inconsistency resolution for knowledge graph generation rules. 10(6):1071–1086. https://doi.org/10.3233/SW-190358.

Heyvaert, P., Dimou, A., Herregodts, A.-L., Verborgh, R., Schuurman, D., Mannens, E., and Van de Walle, R. (2016). RMLEditor: A graph-based mapping editor for linked data mappings. In Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S. P., and Lange, C., editors, *The Semantic Web. Latest Advances and New Domains*, Lecture Notes in Computer Science, pages 709–723. Springer International Publishing. https://doi.org/10.1007/978-3-319-34129-3_43.

HIMMS (2020). Interoperability in healthcare. https://www.himss.org/resources/interoperability-healthcare.

HL7 (2022). Hl7 fhir release v4.3.0. Fast healthcare interoperability resources, HL7. http://hl7.org/fhir/.

Hogan, A. (2020). The semantic web: Two decades on. 11(1):169–185. https://doi.org/10.3233/SW-190387.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2022). Knowledge graphs. 54(4):1–37. https://doi.org/10.1145/3447772.

Holman, C. D. J., Bass, J. A., Rosman, D. L., Smith, M. B., Semmens, J. B., Glasson, E. J., Brook, E. L., Trutwein, B., Rouse, I. L., Watson, C. R., Klerk, N. H. d., and Stanley, F. J. (2008). A decade of data linkage in western australia: strategic design, applications and benefits of the WA data linkage system. 32(4):766–777. https://doi.org/10.1071/ah080766.

Hooper, L. G. and Kaufman, J. D. (2018). Ambient air pollution and clinical implications for susceptible populations. 15:S64–S68. https://doi.org/10.1513/AnnalsATS.201707-574MG.

Hu, W., Qiu, H., Huang, J., and Dumontier, M. (2017). BioSearch: a semantic search engine for Bio2RDF. *Database*, 2017. https://doi.org/10.1093/database/bax059.

Iannella, R. and Villata, S. (2018). Odrl information model 2.2. W3c recommendation, W3C. https://www.w3.org/TR/odrl-model/.

ICO (2012). Anonymisation: managing data protection risk code of practice. Technical report, Information Commisioner's Office. https://ico.org.uk/media/1061/anonymisation-code.pdf.

IDLab (2017). Turtle validator. http://ttl.summerofcode.be/.

Institute, F. C. (2022). Scientists reveal how air pollution can cause lung cancer in people who have never smoked. https://www.crick.ac.uk/news/2022-09-10_scientists-reveal-how-air-pollution-can-cause-lung-cancer-in-people-who-have-never-smoked.

ISGlobal (2023). Climate and health research group at ISGlobal. https://www.isglobal.org/en/-/clima-y-salud.

Ives, C., Pan, H., Edwards, S. W., Nelms, M., Covert, H., Lichtveld, M. Y., Harville, E. W., Wickliffe, J. K., Zijlmans, W., and Hamilton, C. M. (2022). Linking complex disease and exposure data—insights from an environmental and occupational health study. pages 1–5. https://doi.org/10.1038/s41370-022-00428-7.

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W., Imming, M., Jeffery, K. G., Kaliyaperumal, R., Kersloot, M. G., Kirkpatrick, C. R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., da Silva Santos, L. O. B., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M. D., Willighagen, E. L., Wittenburg, P., Roos, M., Mons, B., and Schultes, E. (2020a). FAIR principles: Interpretations and implementation considerations. 2(1):10–29. https://doi.org/10.1162/dint_r_00024.

Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., and Thompson, M. (2020b). A generic workflow for the data FAIRification process. 2(1):56–65. https://doi.org/10.1162/dint_a_00028.

Janowicz, K., Hitzler, P., Li, W., Rehberger, D., Schildhauer, M., Zhu, R., Shimizu, C., Fisher, C. K., Cai, L., Mai, G., Zalewski, J., Zhou, L., Stephen, S., Gonzalez, S., Mecum, B., Lopez-Carr, A., Schroeder, A., Smith, D., Wright, D., Wang, S., Tian, Y., Liu, Z., Shi, M., D'Onofrio, A., Gu, Z., and Currier, K. (2022). Know, know where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence. 43(1):30–39. https://doi.org/10.1002/aaai.12043.

Jennette, J. C., Falk, R. J., Bacon, P. A., Basu, N., Cid, M. C., Ferrario, F., Flores-Suarez, L. F., Gross, W. L., Guillevin, L., Hagen, E. C., Hoffman, G. S., Jayne, D. R., Kallenberg, C. G. M., Lamprecht, P., Langford, C. A., Luqmani, R. A., Mahr, A. D., Matteson, E. L., Merkel, P. A., Ozen, S., Pusey, C. D., Rasmussen, N., Rees, A. J., Scott, D. G. I., Specks, U., Stone, J. H., Takahashi, K., and Watts, R. A. (2013). 2012 revised international chapel hill consensus conference nomenclature of vasculitides. 65(1):1–11. https://doi.org/10.1002/art.37715.

Joinup (2023). Generic RDF validator | joinup. https://joinup.ec.europa.eu/collection/interoperability-test-bed-repository/solution/rdf-validator/generic-rdf-validator.

Jokela, T., Iivari, N., Matero, J., and Karukka, M. (2003). The standard of user-centered design and the standard definition of usability: analyzing ISO 13407 against ISO 9241-11. In *Proceedings of the Latin American conference on Human-computer interaction*, CLIHC '03. Association for Computing Machinery. https://doi.org/10.1145/944519.944525.

Kamdar, M. R., Fernández, J. D., Polleres, A., Tudorache, T., and Musen, M. A. (2019). Enabling web-scale data integration in biomedicine through linked open data. 2(1):90. https://doi.org/10.1038/s41746-019-0162-5.

Kamm, S., Jazdi, N., and Weyrich, M. (2021). Knowledge discovery in heterogeneous and unstructured data of industry 4.0 systems: Challenges and approaches. 104:975–980. https://doi.org/10.1016/j.procir.2021.11.164.

Khalili, A. and Meroño-Peñuela, A. (2017). WYSIWYQ — what you see is what you query. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Third International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 1947 of *CEUR Workshop Proceedings*, pages 123–130. CEUR. https://ceur-ws.org/Vol-1947/#paper11.

Kim, D., Chen, Z., Zhou, L.-F., and Huang, S.-X. (2018). Air pollutants and early origins of respiratory diseases. 4(2):75–94. https://doi.org/10.1016/j.cdtm.2018.03.003.

Kim, G. B. (2019). Reality of kawasaki disease epidemiology. 62(8):292–296. https://doi.org/10.3345/kjp.2019.00157.

Kitching, A. R., Anders, H.-J., Basu, N., Brouwer, E., Gordon, J., Jayne, D. R., Kullman, J., Lyons, P. A., Merkel, P. A., Savage, C. O. S., Specks, U., and Kain, R. (2020). ANCA-associated vasculitis. *Nature Reviews Disease Primers*, 6(1). https://doi.org/10.1038/s41572-020-0204-y.

Klímek, J., Škoda, P., and Nečaský, M. (2019). Survey of tools for linked data consumption. 10(4):665–720. https://doi.org/10.3233/SW-180316.

Knowledge, H. (2023). Health knowledge: health event data. https://www.healthknowledge.org.uk/e-learning/health-information/population-health-practitioners/health-event-data.

Knublauch, H. and Kontokostas, D. (2017). Shapes constraint language (SHACL). W3c recommendation, W3C. https://www.w3.org/TR/shacl/.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., and Zaveri, A. (2014). Test-driven evaluation of linked data quality. In *Proceedings of the 23rd international conference on World wide web - WWW '14*, pages 747–758. ACM Press. https://doi.org/10.1145/2566486.2568002.

Krishnakumar, K. (2002). Intelligent systems for aerospace engineering: An overview. Von Karman Institute Lecture Series on Intelligent Systems for Aeronautics. NASA Technical Reports. https://ntrs.nasa.gov/citations/20020065377.

Kulahcioglu, T., Fradkin, D., Parlak, A., and Belkov, A. (2020). LogVis: Graph-assisted visual analysis of event logs from industrial equipment. In Ivanova, V., Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 2778 of *CEUR Workshop Proceedings*, pages 61–72. CEUR. https://ceur-ws.org/Vol-2778/#paper6.

Kuric, E., Fernández, J. D., and Drozd, O. (2019). Knowledge graph exploration: A usability evaluation of query builders for laypeople. In Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., and Sure-Vetter, Y., editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, International Conference on Semantic Systems, pages 326–342. Springer International Publishing. https://doi.org/10.1007/978-3-030-33220-4_24.

Křemen, P., Saeeda, L., Blaško, M., and Med, M. (2018). Dataset dashboard - a SPARQL endpoint explorer. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Fourth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 2187 of *CEUR Workshop Proceedings*, pages 70–77. CEUR. https://ceur-ws.org/Vol-2187/#paper7.

Lebo, T., Sahoo, S. S., and McGuinness, D. L. (2013). PROV-O: The PROV Ontology. W3c recommendation, W3C. https://www.w3.org/TR/prov-o/.

Lefort, L., Haller, A., Taylor, K., Squire, G., Taylor, P., Percival, D., and Woolf, A. (2017). The ACORN-SAT linked climate dataset. 8(6):959–967. https://doi.org/10.3233/SW-160241.

Leskinen, P., Miyakita, G., Koho, M., and Hyvönen, E. (2018). Combining faceted search with data-analytic visualizations on top of a SPARQL endpoint. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Fourth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 2187 of *CEUR Workshop Proceedings*, pages 53–63. CEUR. https://ceur-ws.org/Vol-2187/#paper5.

Lewis, J. (2002). Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human–Computer Interaction*, 14(3-4):463–488. https://doi.org/10.1080/10447318.2002.9669130.

Lewis, J. R. (1992). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. 36(16):1259–1260. https://doi.org/10.1177/154193129203601617.

Lewis, J. R. (2014). Usability: Lessons learned ... and yet to be learned. 30(9):663–684. https://doi.org/10.1080/10447318.2014.930311.

Lochmüller, H., Badowska, D. M., Thompson, R., Knoers, N. V., Aartsma-Rus, A., Gut, I., Wood, L., Harmuth, T., Durudas, A., Graessner, H., Schaefer, F., and

Riess, O. (2018). RD-connect, NeurOmics and EURenOmics: collaborative european initiative for rare diseases. 26(6):778–785. https://doi.org/10.1038/s41431-018-0115-5.

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P., and Reichstein, M. (2020). Earth system data cubes unravel global multivariate dynamics. 11(1):201–234. https://doi.org/10.5194/esd-11-201-2020.

Maitre, L., Guimbaud, J.-B., Warembourg, C., Güil-Oumrait, N., Petrone, P. M., Chadeau-Hyam, M., Vrijheid, M., Basagaña, X., and Gonzalez, J. R. (2022). State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event. 168:107422. https://doi.org/10.1016/j.envint.2022.107422.

Makino, N., Nakamura, Y., Yashiro, M., Kosami, K., Matsubara, Y., Ae, R., Aoyama, Y., and Yanagawa, H. (2019). Nationwide epidemiologic survey of kawasaki disease in japan, 2015–2016. 61(4):397–403. https://doi.org/10.1111/ped.13809.

Maramba, I., Chatterjee, A., and Newman, C. (2019). Methods of usability testing in the development of eHealth applications: A scoping review. 126:95–104. https://doi.org/10.1016/j.ijmedinf.2019.03.018.

Marcilly, R., Douze, L., Ferré, S., Audeh, B., Bobed, C., Lillo-Le Louët, A., Lamy, J.-B., and Bousquet, C. (2020). How to interact with medical terminologies? Formative usability evaluations comparing three approaches for supporting the use of MedDRA by pharmacovigilance specialists. *BMC Medical Informatics and Decision Making*, 20(1):261. https://doi.org/10.1186/s12911-020-01280-1.

Martín Salas, J. and Harth, A. (2012). NeoGeo vocabulary specification. Technical report, GeoVocab. http://geovocab.org/doc/neogeo/.

MET, (2023). Climate historical data from met Éireann. https://www.met.ie//climate/available-data/historical-data.

Mina, E., Thompson, M., Hettne, K. M., Roon-Mom, W. V., Kaliyaperumal, R., Horst, E. V. D., Wolstencroft, K., Mons, B., and Roos, M. (2015). Multidisciplinary collaboration to facilitate hypotheses generation in huntington's disease. In *2015 IEEE 11th International Conference on e-Science*, pages 118–125. https://doi.org/10.1109/eScience.2015.71.

Minutolo, A., Damiano, E., De Pietro, G., Fujita, H., and Esposito, M. (2022). A conversational agent for querying italian patient information leaflets and improving health literacy. *Computers in Biology and Medicine*, 141:105004. https://doi.org/10.1016/j.compbiomed.2021.105004.

Mitschick, A., Nieschalk, F., Voigt, M., and Dachselt, R. (2017). IcicleQuery: A web search interface for fluid semantic query construction. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Third International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 1947 of *CEUR Workshop Proceedings*, pages 99–110. CEUR. `https://ceur-ws.org/Vol-1947/#paper09`.

Murrison, L. B., Brandt, E. B., Myers, J. B., and Hershey, G. K. K. (2019). Environmental exposures and mechanisms in allergy and asthma development. *The Journal of Clinical Investigation*, 129(4):1504–1515. `https://doi.org/10.1172/JCI124612`.

Navarro-Gallinad, A. (2021). Weather and air quality data for ireland as RDF data cube. Technical report. `https://doi.org/10.5281/zenodo.5668286`.

Navarro-Gallinad, A. (2022). NUTS-RDF in GeoSPARQL. Technical report. `https://doi.org/10.5281/zenodo.6514296`.

Navarro-Gallinad, A. (2023). serdif. Technical report. `https://w3id.org/serdif`.

Navarro-Gallinad, A., Meehan, A., and O'Sullivan, D. (2020). The Semantic Combining for Exploration of Environmental and Disease Data Dashboard for Clinician Researchers. In Ivanova, V., Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 2778 of *VOILA'20*, pages 73–85, Virtual Workshop. CEUR Workshop Proceedings. `http://ceur-ws.org/Vol-2778/paper7.pdf`.

Navarro-Gallinad, A., Orlandi, F., and O'Sullivan, D. (2021). Enhancing Rare Disease Research with Semantic Integration of Environmental and Health Data. In *The 10th International Joint Conference on Knowledge Graphs*, IJCKG'21, pages 19–27, New York, NY, USA. Association for Computing Machinery. `https://doi.org/10.1145/3502223.3502226`.

Navarro-Gallinad, A., Orlandi, F., and O'Sullivan, D. (2023). Environmental data associated to particular health events example dataset. Technical report. `https://doi.org/10.5281/zenodo.5544257`.

Navarro-Gallinad, A., Orlandi, F., Scott, J., Little, M., and O'Sullivan, D. (2022). Evaluating the usability of a semantic environmental health data framework: Approach and study. Preprint:1–24. `https://doi.org/10.3233/SW-223212`.

Nikolaou, C., Dogani, K., Bereta, K., Garbis, G., Karpathiotakis, M., Kyzirakos, K., and Koubarakis, M. (2015). Sextant: Visualizing time-evolving linked geospatial data. 35:35–52. `https://doi.org/10.1016/j.websem.2015.09.004`.

Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. 16(1):160940691773384. `https://doi.org/10.1177/1609406917733847`.

Oldman, D. and Tanase, D. (2018). Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace. In Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M. C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A., and Simperl, E., editors, *The Semantic Web – ISWC 2018*, Lecture Notes in Computer Science, pages 325–340. Springer International Publishing. `https://doi.org/10.1007/978-3-030-00668-6_20`.

Ontotext (2022). GraphDB downloads and resources. `http://graphdb.ontotext.com/`.

OSI (2023). Geohive. `https://www.geohive.ie/`.

Oxford, A. L. D. (2023). Oxford advanced learner's dictionary: Effective adjective. `https://www.oxfordlearnersdictionaries.com/definition/english/effective?q=effective`.

Pandit, H. J. (2022). Data privacy vocabulary (DPV). Final community group report, W3C. `https://w3id.org/dpv`.

Pandit, H. J., Polleres, A., Bos, B., Brennan, R., Bruegger, B., Ekaputra, F. J., Fernández, J. D., Hamed, R. G., Kiesling, E., Lizar, M., Schlehahn, E., Steyskal, S., and Wenning, R. (2019). Creating a vocabulary for data privacy. In Panetto, H., Debruyne, C., Hepp, M., Lewis, D., Ardagna, C. A., and Meersman, R., editors, *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, Lecture Notes in Computer Science, pages 714–730. Springer International Publishing. `https://doi.org/10.1007/978-3-030-33246-4_44`.

Parra, G. d. l. and Hogan, A. (2021). Fast approximate autocompletion for SPARQL query builders. In Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Sixth International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3023 of *CEUR Workshop Proceedings*, pages 41–55. CEUR. `https://ceur-ws.org/Vol-3023/#paper10`.

Pastorino, R., De Vito, C., Migliara, G., Glocker, K., Binenbaum, I., Ricciardi, W., and Boccia, S. (2019). Benefits and challenges of big data in healthcare: an overview of the european initiatives. 29:23–27. `https://doi.org/10.1093/eurpub/ckz168`.

Patella, V., Florio, G., Magliacane, D., Giuliano, A., Crivellaro, M. A., Di Bartolomeo, D., Genovese, A., Palmieri, M., Postiglione, A., Ridolo, E., Scaletti, C., Ventura, M. T., Zollo, A., Pollution, A., Climate Change Task Force of the Italian Society of Allergology, A., and (SIAAIC), C. I. (2018). Urban air pollution and climate change: "the decalogue: Allergy safe tree" for allergic and respiratory diseases care. 16(1):20. `https://doi.org/10.1186/s12948-018-0098-3`.

Perry, M., Herring, J., Car, N. J., Homburg, T., Abhayaratna, J., Cox, S. J. D., Bonduel, M., and Knibbe, F. (2012). GeoSPARQL - A Geographic Query Language for RDF Data | OGC. Ogc standard, Open Geospatial Consortium. `https://www.ogc.org/standards/geosparql`.

Pesquita, C., Ivanova, V., Lohmann, S., and Lambrix, P. (2018). A framework to conduct and report on empirical user studies in semantic web contexts. In Faron Zucker, C., Ghidini, C., Napoli, A., and Toussaint, Y., editors, *Knowledge Engineering and Knowledge Management*, European Knowledge Acquisition Workshop, pages 567–583. Springer International Publishing. `https://doi.org/10.1007/978-3-030-03667-6_36`.

Peterson, D., Gao, S., Malhotra, A., Sperberg-McQueen, C. M., and Thompson, H. S. (2012). W3c XML schema definition language (XSD) 1.1 part 2: Datatypes. W3c recommendation, W3C. `https://www.w3.org/TR/xmlschema11-2/`.

Piel, F. B., Fecht, D., Hodgson, S., Blangiardo, M., Toledano, M., Hansell, A. L., and Elliott, P. (2020). Small-area methods for investigation of environment and health. 49(2):686–699. `https://doi.org/10.1093/ije/dyaa006`.

Pinero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L. I. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. 2015(0):bav028–bav028. `https://doi.org/10.1093/database/bav028`.

Plotly (2023). Dash documentation & user guide | plotly. `https://dash.plotly.com/`.

Prud'hommeaux, E. (2006). W3c RDF validation service. Technical report, W3C. `https://www.w3.org/RDF/Validator/`.

Raissya, H., Darari, F., and Ekaputra, F. J. (2021). VizKG: A framework for visualizing SPARQL query results over knowledge graphs (short paper). In Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Sixth International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3023 of *CEUR Workshop Proceedings*, pages 95–102. CEUR. `https://ceur-ws.org/Vol-3023/#paper3`.

Rajagopalan, S., Al, K. S. G., and Brook, R. D. (2018). Air pollution and cardiovascular disease. 72(17):2054–2070. `https://doi.org/10.1016/j.jacc.2018.07.099`.

Ramachandran, R., Bugbee, K., and Murphy, K. (2021). From open data to open science. 8(5):e2020EA001562. `https://doi.org/10.1029/2020EA001562`.

Ramadhana, N. H., Darari, F., Putra, P. O. H., Nutt, W., Razniewski, S., and Akbar, R. I. (2020). User-centered design for knowledge imbalance analysis: A case study of ProWD. In Ivanova, V., Lambrix, P., Pesquita, C., and Wiens, V., editors, *Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 2778 of *CEUR Workshop Proceedings*, pages 14–27. CEUR. `https://ceur-ws.org/Vol-2778/#paper2`.

Rampin, R. and Rampin, V. (2021). Taguette: open-source qualitative data analysis. 6(68):3522. `https://doi.org/10.21105/joss.03522`.

Ramírez-Andreotta, M. D., Walls, R., Youens-Clark, K., Blumberg, K., Isaacs, K. E., Kaufmann, D., and Maier, R. M. (2021). Alleviating environmental health disparities through community science and data integration. 5. https://doi.org/10.3389/fsufs.2021.620470.

Randles, A. and O'Sullivan, D. (2022). Evaluating quality improvement techniques within the linked data generation process. pages 21–35. https://doi.org/10.3233/SSW220006.

Reddy, B. P., Houlding, B., Hederman, L., Canney, M., Debruyne, C., O'Brien, C., Meehan, A., O'Sullivan, D., and Little, M. A. (2019). Data linkage in medical science using the resource description framework: the AVERT model. 1:20. https://doi.org/10.12688/hrbopenres.12851.2.

Regalia, B., Janowicz, K., and Mai, G. (2017). Phuzzy.link: A SPARQL-powered client-sided extensible semantic web browser. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Third International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 1947 of *CEUR Workshop Proceedings*, pages 34–44. CEUR. https://ceur-ws.org/Vol-1947/#paper04.

Rietveld, L. and Hoekstra, R. (2017). The YASGUI family of SPARQL clients 1. 8(3):373–383. https://doi.org/10.3233/SW-150197.

Rife, E. and Gedalia, A. (2020). Kawasaki disease: an update. 22(10):75. https://doi.org/10.1007/s11926-020-00941-4.

Rodó, X., Ballester, J., Curcoll, R., Boyard-Micheau, J., Borràs, S., and Morguí, J.-A. (2016). Revisiting the role of environmental and climate factors on the epidemiology of Kawasaki disease. *Annals of the New York Academy of Sciences*, 1382(1):84–98. https://doi.org/10.1111/nyas.13201.

Rodó, X., Curcoll, R., Robinson, M., Ballester, J., Burns, J. C., Cayan, D. R., Lipkin, W. I., Williams, B. L., Couto-Rodriguez, M., Nakamura, Y., Uehara, R., Tanimoto, H., and Morguí, J.-A. (2014). Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. *Proceedings of the National Academy of Sciences*, 111(22):7952–7957. https://doi.org/10.1073/pnas.1400380111.

Roos, A., Beltran, S., Piscia, D., Laurie, S., Protasio, J., Cañada, A., Fernández, J., Kaliyaperumal, R., Lair, S., Sernadela, P., Girdea, M., Thompson, R., Straub, V., Roos, M., T'Hoen, P., Valencia, A., Salgado, D., Béroud, C., Gut, I., and Lochmüller, H. (2016). RD-connect: Data sharing and analysis for rare disease research within the integrated platform and through GA4gh beacon and matchmaker exchange. 26:S160–S161. https://doi.org/10.1016/j.nmd.2016.06.272.

Roos, M., López Martin, E., and Wilkinson, M. D. (2017). Preparing data at the source to foster interoperability across rare disease resources. In Posada de la Paz, M., Taruscio, D., and Groft, S. C., editors, *Rare Diseases Epidemiology:*

Update and Overview, Advances in Experimental Medicine and Biology, pages 165–179. Springer International Publishing. https://doi.org/10.1007/978-3-319-67144-4_9.

Roussey, C., Bernard, S., André, G., and Boffety, D. (2020). Weather data publication on the LOD using SOSA/SSN ontology. 11(4):581–591. https://doi.org/10.3233/SW-200375.

Rowley, A. H. and Shulman, S. T. (2018). The epidemiology and pathogenesis of kawasaki disease. 6. https://doi.org/10.3389/fped.2018.00374.

Sauro, J. and Lewis, J. R. (2016). *Quantifying the user experience: practical statistics for user research*. Morgan Kaufmann, 2nd edition edition.

Schaaf, J., Sedlmayr, M., Schaefer, J., and Storf, H. (2020). Diagnosis of rare diseases: a scoping review of clinical decision support systems. 15(1):263. https://doi.org/10.1186/s13023-020-01536-z.

Schraufnagel, D. E., Balmes, J. R., Cowl, C. T., De Matteis, S., Jung, S.-H., Mortimer, K., Perez-Padilla, R., Rice, M. B., Riojas-Rodriguez, H., Sood, A., Thurston, G. D., To, T., Vanker, A., and Wuebbles, D. J. (2019). Air pollution and noncommunicable diseases. 155(2):417–426. https://doi.org/10.1016/j.chest.2018.10.041.

Scott, J., Hartnett, J., Mockler, D., and Little, M. A. (2020). Environmental risk factors associated with ANCA associated vasculitis: A systematic mapping review. 19(11):102660. https://doi.org/10.1016/j.autrev.2020.102660.

Sernadela, P., González-Castro, L., Carta, C., van der Horst, E., Lopes, P., Kaliyaperumal, R., Thompson, M., Thompson, R., Queralt-Rosinach, N., Lopez, E., Wood, L., Robertson, A., Lamanna, C., Gilling, M., Orth, M., Merino-Martinez, R., Posada, M., Taruscio, D., Lochmüller, H., Robinson, P., Roos, M., and Oliveira, J. L. (2017a). Linked registries: Connecting rare diseases patient registries through a semantic web layer. 2017:e8327980. https://doi.org/10.1155/2017/8327980.

Sernadela, P., González-Castro, L., and Oliveira, J. L. (2017b). SCALEUS: Semantic web services integration for biomedical applications. 41(4):54. https://doi.org/10.1007/s10916-017-0705-8.

Serrano, F., Nunes, S., and Pesquita, C. (2022). VOWLExplain: Knowledge graph visualization for explainable artificial intelligence. In Fu, B., Lambrix, P., and Pesquita, C., editors, *Proceedings of the Seventh International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3253 of *CEUR Workshop Proceedings*, pages 29–40. CEUR. https://ceur-ws.org/Vol-3253/#paper3.

Sillé, F. C. M., Karakitsios, S., Kleensang, A., Koehler, K., Maertens, A., Miller, G. W., Prasse, C., Quiros-Alcala, L., Ramachandran, G., and Hartung, T. (2020). The exposome : a new approach for risk assessment. 37(1):3–23. https://doi.org/10.14573/altex.2001051.

Soylu, A., Kharlamov, E., Zheleznyakov, D., Jimenez-Ruiz, E., Giese, M., and Horrocks, I. (2015). OptiqueVQS: Ontology-based visual querying. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the International Workshop on Visualizations and User Interfaces for Ontologies and Linked Data*, volume 1456 of *CEUR Workshop Proceedings*, page 91. CEUR. https://ceur-ws.org/Vol-1456/#paper10.

Standing Committee on Emerging Science for Environmental Health Decisions, Board on Life Sciences, Board on Environmental Studies and Toxicology, Division on Earth and Life Studies, and National Academies of Sciences, Engineering, and Medicine (2018). *Informing Environmental Health Decisions Through Data Integration: Proceedings of a Workshop in Brief.* National Academies Press. https://doi.org/10.17226/25139.

Stöhr, M. R., Günther, A., and Majeed, R. W. (2021). The Collaborative Metadata Repository (CoMetaR) Web App: Quantitative and Qualitative Usability Evaluation. *Journal of Medical Internet Research Medical Informatics*, 9(11):e30308. https://doi.org/10.2196/30308.

Tartari, G. and Hogan, A. (2018). WiSP: Weighted shortest paths for RDF graphs. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Fourth International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 2187 of *CEUR Workshop Proceedings*, pages 37–52. CEUR. https://ceur-ws.org/Vol-2187/#paper4.

Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Béroud, C., Gut, I. G., Hansson, M. G., Hoen, P.-B. A. Patrinos, G. P., Dawkins, H., Ensini, M., Zatloukal, K., Koubi, D., Heslop, E., Paschall, J. E., Posada, M., Robinson, P. N., Bushby, K., and Lochmüller, H. (2014). RD-connect: An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. 29(3):780–787. https://doi.org/10.1007/s11606-014-2908-8.

Trinh, T.-D., Wetz, P., Do, B.-L., Aryan, P. R., Kiesling, E., and Tjoa, A. M. (2015). An autocomplete input box for semantic annotation on the web. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the International Workshop on Visualizations and User Interfaces for Ontologies and Linked Data*, volume 1456 of *CEUR Workshop Proceedings*, page 97. CEUR. https://ceur-ws.org/Vol-1456/#paper11.

Turner, M. C., Andersen, Z. J., Baccarelli, A., Diver, W. R., Gapstur, S. M., Pope III, C. A., Prada, D., Samet, J., Thurston, G., and Cohen, A. (2020). Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. 70(6):460–479. https://doi.org/10.3322/caac.21632.

Uehara, R. and Belay, E. D. (2012). Epidemiology of kawasaki disease in asia, europe, and the united states. 22(2):79–85. https://doi.org/10.2188/jea.JE20110131.

UK, H. (2023). Health data research united kingdom (HDR UK): Health data research. https://www.hdruk.ac.uk/about-us/what-we-do/health-data-research-explained/.

Ulrich, H., Kock-Schoppenhauer, A.-K., Deppenwiese, N., Gött, R., Kern, J., Lablans, M., Majeed, R. W., Stöhr, M. R., Stausberg, J., Varghese, J., Dugas, M., and Ingenerf, J. (2022). Understanding the nature of metadata: Systematic review. 24(1):e25440. https://doi.org/10.2196/25440.

UN (2022). United nations news: Un general assembly declares access to clean and healthy environment a universal human right. https://news.un.org/en/story/2022/07/1123482.

Vega-Gorgojo, G., Giménez-García, J. M., Ordóñez, C., and Bravo, F. (2022). Pioneering easy-to-use forestry data with forest explorer. 13(2):147–162. https://doi.org/10.3233/SW-210430.

Vega-Gorgojo, G., Slaughter, L., Giese, M., Heggestøyl, S., Soylu, A., and Waaler, A. (2016). Visual query interfaces for semantic datasets: An evaluation study. 39:81–96. https://doi.org/10.1016/j.websem.2016.01.002.

W3C (2015). Semantic web. W3c recommendation. https://www.w3.org/standards/semanticweb/.

W3C SPARQL, W. G. (2013). SPARQL Query Language for RDF. W3C recommendation. W3c recommendation. http://www.w3.org/TR/sparql11-overview/.

Wang, J., Wang, X., Ma, C., and Kou, L. (2021). A survey on the development status and application prospects of knowledge graph in smart grids. 15(3):383–407. https://doi.org/10.1049/gtd2.12040.

Weise, M., Lohmann, S., and Haag, F. (2016). LD-VOWL: Extracting and visualizing schema information for linked data endpoints. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Second International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 1704 of *CEUR Workshop Proceedings*, pages 120–127. CEUR. https://ceur-ws.org/Vol-1704/#paper11.

WHO (2023a). World health organization: Air pollution. https://www.who.int/health-topics/air-pollution#tab=tab_1.

WHO (2023b). World health organization: Environmental health. https://www.who.int/health-topics/environmental-health.

Wilcke, W. X., de Boer, V., de Kleijn, M. T. M., van Harmelen, F. A. H., and Scholten, H. J. (2019). User-centric pattern mining on knowledge graphs: An archaeological case study. 59:100486. https://doi.org/10.1016/j.websem.2018.12.004.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., ’t Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018. `https://doi.org/10.1038/sdata.2016.18`.

WordReference (2023). Word reference dictionary of english: event. `https://www.wordreference.com/definition/event`.

Xu, X., Nie, S., Ding, H., and Hou, F. F. (2021). Environmental pollution and kidney diseases. 14(5):313–324. `https://doi.org/10.1038/nrneph.2018.11`.

Yacoubi, N., Graux, D., and Faron, C. (2022). Multi-level visual tours of weather linked data. In Fu, B., Lambrix, P., and Pesquita, C., editors, *Proceedings of the Seventh International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3253 of *CEUR Workshop Proceedings*, pages 52–57. CEUR. `https://ceur-ws.org/Vol-3253/#paper5`.

Zainab, S. S. e., Saleem, M., Mehmood, Q., Zehra, D., Decker, S., and Hasnain, A. (2015). FedViz: A visual interface for SPARQL queries formulation and execution. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the International Workshop on Visualizations and User Interfaces for Ontologies and Linked Data*, volume 1456 of *CEUR Workshop Proceedings*, page 49. CEUR. `https://ceur-ws.org/Vol-1456/#paper5`.

Zaitchik, B. F., Sweijd, N., Shumake-Guillemot, J., Morse, A., Gordon, C., Marty, A., Trtanj, J., Luterbacher, J., Botai, J., Behera, S., Lu, Y., Olwoch, J., Takahashi, K., Stowell, J. D., and Rodó, X. (2020). A framework for research linking weather, climate and COVID-19. 11(1):5730. `https://doi.org/10.1038/s41467-020-19546-7`.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. 7(1):63–93. `https://doi.org/10.3233/SW-150175`.

Čerāns, K., Bārzdiņš, J., Šostaks, A., Ovčiņņikova, J., Lāce, L., Grasmanis, M., and Sproģis, A. (2017). Extended UML class diagram constructs for visual SPARQL queries in ViziQuer/web. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Third International Workshop on Visualization and*

Interaction for Ontologies and Linked Data, volume 1947 of *CEUR Workshop Proceedings*, pages 87–98. CEUR. `https://ceur-ws.org/Vol-1947/#paper08`.

Čerāns, K. and Ovčiņņikova, J. (2016). ViziQuer: Notation and tool for data analysis SPARQL queries. In Ivanova, V., Lambrix, P., Lohmann, S., and Pesquita, C., editors, *Proceedings of the Second International Workshop on Visualization and Interaction for Ontologies and Linked Data*, volume 1704 of *CEUR Workshop Proceedings*, pages 151–159. CEUR. `https://ceur-ws.org/Vol-1704/#paper15`.

Čerāns, K., Ovčiņņikova, J., Grasmanis, M., and Lāce, L. (2022). Experience with visual SPARQL queries over DBPedia. In Fu, B., Lambrix, P., and Pesquita, C., editors, *Proceedings of the Seventh International Workshop on the Visualization and Interaction for Ontologies and Linked Data*, volume 3253 of *CEUR Workshop Proceedings*, pages 59–65. CEUR. `https://ceur-ws.org/Vol-3253/#paper6`.

A SERDIF R2RML mapping example

Listing A.1: SERDIF R2RML mapping example.

```
# -- 1. Namespaces -----------------------------------
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rrf: <http://kdeg.scss.tcd.ie/ns/rrf#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> .
# -- 2. Triple map for observation data -------------------------
<#MapObsData>
# -- 3. Select dates from datetimes ------------------------------
   rr:logicalTable [
    rr:sqlQuery """
 SELECT CAST(DATETIMEBEGIN AS varchar(10)) AS TIMEG, AIRPOLLUTANT,
 CAST(AVG(CAST(CONCENTRATION AS FLOAT)) AS DECIMAL (10,2)) AS CONCENTRATION
 FROM {{data.eeaDataFile}}
 GROUP BY TIMEG
     """;
   ] ;
# -- 4. Define observation (subject) based on lat, lon and date --
   rr:subjectMap [
    rr:template "https://serdif.adaptcentre.ie/kg/2022/dataset#type=
       airquality&source=eea&version=vE1a&point={{data.lon}}_{{data.lat}}&
       time={TIMEG}";
    rr:termType rr:IRI;
    rr:class qb:Observation;
    rr:graphMap [ rr:template "https://serdif.adaptcentre.ie/kg/2022/dataset
       #type=airquality&source=eea&version=vE1a&point={{data.lon}}_{{data.
       lat}}" ] ;
   ];
# -- 5. Link observations with dataset through location ----------
   rr:predicateObjectMap [
    rr:predicate qb:dataSet;
    rr:objectMap [
      rr:template "https://serdif.adaptcentre.ie/kg/2022/dataset#type=
         airquality&source=eea&version=vE1a&point={{data.lon}}_{{data.lat}}
         ";
      rr:termType rr:IRI;
    ];
   ];
# -- 6. Define time dimension for each observaion ----------------
   rr:predicateObjectMap [
    rr:predicate sdmx-dimension:timePeriod ;
    rr:objectMap [
    rrf:functionCall [
           rrf:function <#time2datetime> ;
           rrf:parameterBindings (
             [ rr:column "TIMEG" ; ]
           ) ;
         ] ;
    rr:termType rr:Literal;
    rr:datatype xsd:dateTime;
    ];
  ];
# -- 7. Include air pollutant concentration values ---------------
   rr:predicateObjectMap [
    rr:predicateMap [
      rrf:functionCall [
           rrf:function <#pollutantNameClean> ;
           rrf:parameterBindings (
             [ rr:column "AIRPOLLUTANT" ]
           ) ;
         ] ;
      rr:termType rr:IRI;
    ];
    rr:objectMap [
      rr:column "CONCENTRATION";
      rr:termType rr:Literal;
      rr:datatype xsd:float;
    ];
   ];
.
# -- 8. Define a function to format datetime values ------------
<#time2datetime>
   rrf:functionName "time2datetime" ;
   rrf:functionBody """
    function time2datetime(timeC) {
    // From 2010-01-01 to 2010-01-01T00:00:00
    return String(timeC ) + "T00:00:00" ;
   }
```

```
    """ ;
.
# -- 9. Define a function to clean pollutant names -------------
<#pollutantNameClean>
    rrf:functionName "pollutantnameclean" ;
    rrf:functionBody """
     function pollutantnameclean(pName) {
     // Fix format pollutant name to comply with URI standard symbols
     // by replacing parentheses, dashes, plus signs and commasdate time to
         conform with standards
     // "Indeno-(1,2,3-cd)pyrene in PM10" -> "Indeno123cdpyreneinPM10"
     var pNameC = pName.replace(/[{()}]/g, ’’).replace(/\\-/g, ’’).replace
         (/\\+/g, ’’).replace(/,/g, ’’).replace(/\\s/g, ’’).replace(/\\=/g,
         ’’) ;
     return "https://serdif.adaptcentre.ie/kg/2022/measure#has" + pNameC + "
         Value";
    }
    """ ;
.
```

B Informed Consent Form

TRINITY COLLEGE DUBLIN – INFORMED CONSENT FORM

LEAD RESEARCHERS: Albert Navarro Gallinad

BACKGROUND OF RESEARCH: HEalth data LInkage for ClinicAL benefit (HELICAL) is a European project aimed at training PhD students and future researchers in collecting, combining and analysing medical and genetic data. This includes learning how to protect the data on the individual patients so that they cannot be identified. The goal of this project is to find solutions to the challenges faced by patients with rare diseases when it comes to connecting their personal data with scientific data. Our research will lead to understanding of their disease, finding a better treatment and improving their quality of life.

In this project, we aim to address the challenge of integrating multiple heterogeneous data sources using Knowledge Graphs (KG). These technologies have a steep learning curve which can present an obstacle for non-technical researchers who want to access and explore the data to meet their needs. That is why we designed the Semantic Environmental and Rare Disease Integration Framework (SERDIF) User Interface (UI), a visual tool designed for use by Health Data researchers. The SERDIF UI is a visual tool designed to safely combine, access, comprehend and export environmental data associated to individual events through location and time; whilst hiding the complexities of these technologies.

PROCEDURES OF THIS STUDY

– You are going to be briefed on the experiment task, what to do in the experiment.

– You will be asked to use the SERDIF UI and perform some tasks while thinking aloud.

– You will be asked to fill out a usability test survey.

This experiment will take place online over a conference call via a video conferencing platform, with screen sharing and audio enabled from the participant.

The total duration could take up to an hour to perform the tasks and fill the questionnaire. We will track the time you spent in the completion of each task with a stopwatch. While thinking aloud, you will be recorded with an automatic transcription feature of this video conferencing platform. This recording will be used to correct the statements that the note-taker will write down during the experiment. Audio and video will not be recorded during your session. The resulting data stored will be summary tables with the time per task and the numeric answers of the usability test survey, and text files with the open comments of the usability test survey and the automatic transcriptions of the experiment session.

Your data will not be identifiable since it will be coded with a participant number and stored using the IT services called MyZone Google Drive which complies with GDPR rules. Only the lead researcher (Albert Navarro Gallinad) and the two supervisors of the lead researcher (Prof. Declan O'Sullivan, Dr. Fabrizio Orlandi) will have access to this data until its publication in an open data repository.

We will perform a qualitative analysis of the think-aloud data by coding and categorising the statements, once we have the aggregated data from all the participants. Then, resulting emerging themes will be the ones reported as the results of this experiment. Furthermore, the quantitative results from the time per task and the usability test survey will be analysed with statistical summaries, reporting aggregated results.

None of your personal details will be recorded and you are free to stop and leave the experiment at any point if you so choose.

PUBLICATION

The goal will be to publish the results of the usability test at Semantic Web conference and workshops, such as Extended Semantic Web Conference (ESWC) and International Semantic Web Conference (ISWC); and relevant journals such as Web Semantics and the Journal of Biomedical and Health Informatics, as well as the PhD thesis of the lead researcher at Trinity College Dublin.

CONFLICTS OF INTEREST

My supervisors will not take part in the experiment.

Potential participants of the experiment will not be provided with any prior information before the experiment.

Individual results will be anonymized and published in open data repositories for reproducibility and research will be reported on the aggregate results.

DECLARATION

– I am 18 years or older and am competent to provide consent.

– I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.

- I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I understand that if I make illicit activities known, these will be reported to appropriate authorities.
- I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
- I understand that if the results of the research have been published, ¡or my data has been fully anonymised so that it can no longer be attributed to me¿, then it will no longer be possible to withdraw.
- I understand that I may stop electronic recordings at any time, and that I may at any time, even subsequent to my participation [request to] have such recordings destroyed (except in situations such as above).
- I understand that, subject to the constraints above, no recordings will be replayed in any public forum or made available to any audience other than the current researchers/research team.
- I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- I understand that my participation is fully anonymous and that no personal details about me will be recorded.
- I understand that if I or anyone in my family has a history of epilepsy then I am proceeding at my own risk.
- I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation.
- I have received a copy of this agreement.

By signing this document, I consent to participate in this study, and consent to the data processing necessary to enable my participation and to achieve the research goals of this study.

PARTICIPANT'S NAME:

PARTICIPANT'S SIGNATURE:

Date:

Statement of investigator's responsibility: I have explained the nature and purpose of this research study, the procedures to be undertaken and any risks that may be involved. I have offered to answer any questions and fully answered such questions. I believe that the participant understands my explanation and has freely given informed consent.

RESEARCHERS CONTACT DETAILS:

RESEARCHER'S SIGNATURE:

Date:

C PSSUQ questionnaire (v2)

The Post-Study System Usability Questionnaire (PSSUQ)

This questionnaire, which starts on the following page, gives you an opportunity to tell us your reactions to the system you used. Your responses will help us understand what aspects of the system you are particularly concerned about and the aspects that satisfy you.

To as great a degree as possible, think about all the tasks that you have done with the system while you answer these questions.

Please read each statement and indicate how strongly you agree or disagree with the statement by circling a number on the scale. If a statement does not apply to you, circle N/A.

Please write comments to elaborate on your answers.

After you have completed this questionnaire, I'll go over your answers with you to make sure I understand all of your responses.

Thank you!

Each question is optional. Feel free to omit a response to any question; however the researcher would be grateful if all questions are responded to.

1. Overall, I am satisfied with how easy it is to use this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

2. It was simple to use this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

3. I could effectively complete the tasks and scenarios using this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

4. I was able to complete the tasks and scenarios quickly using this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

5. I was able to efficiently complete the tasks and scenarios using this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

6. I felt comfortable using this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

7. It was easy to learn to use this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

8. I believe I could become productive quickly using this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

9. The system gave error messages that clearly told me how to fix problems.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

10. Whenever I made a mistake using the system, I could recover easily and quickly.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

11. The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

12. It was easy to find the information I needed.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

13. The information provided for the system was easy to understand.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

14. The information was effective in helping me complete the tasks and scenarios.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

15. The organization of information on the system screens was clear.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

16. The interface of this system was pleasant.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

17. I liked using the interface of this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

18. This system has all the functions and capabilities I expect it to have.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

19. Overall, I am satisfied with this system.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree
Comments:								

D Code descriptions for the thematic analysis

Table D.1: Description of the codes and references from the participants in the thematic analysis of the usability sessions transcripts (P1).

Code	Description	References
Additional features	Extend data aggregation options. Add more event categories. Custom event input.	15
Complex data linking process	The data linkage process is not clear	31
Complex task instructions	Task wording and structure are difficult to follow	10
Confusing text descriptions	Some text descriptions and dashboard features are not easy to understand	16
Environmental data prior to flare events	Temporal linkage must be for the period before multiple clinical events	23
Good data exploration features	Data exploration features such as data tables, plots and comparing queries engage researchers with the linked data	56
Helpful text, tooltips and summaries	Descriptions, tooltips and pop-ups facilitated the data linkage process and summaries provided a helpful overview of the data for better understanding	24
Moderator assist	The moderator intervenes because the participant struggles to complete a task, wanders off task or goes too deep into a task or the system crashes.	231
Technical session issues	Delayed responses, control malfunctioning and script errors during the session	28
Unclear standardisation of the data	The data standardisation process and the use of z-scores are not clear	9
Useful and easy to use approach	The approach is useful and easy to link environmental and health data for researchers	25

Useful data linkage and export features	Query features are useful to link and retrieve the required linked data	21
Visualization of plots complex	Some visualisations lacked axis labels and introductory text for a better grasp, and grouping queries in a plot was complex	50
Additional features	Extend data aggregation options. Add more event categories. Custom event input.	15

Table D.2: Description of the codes and references from the participants in the thematic analysis of the usability sessions transcripts (P2).

Code	**Description**	**References**
Additional explanation	The participant requests additional explanation for a feature, visualisation, text description or tool tip.	59
Additional feature suggestion	Distinguish outliers based on historical data. Add time series, scatter and histogram plots. Add an advanced aggregation method and select all features. Define, select and group events feature for exploration. Summarise the metadata in a user friendly way. Facilitate the understanding of the data table and increase the amount of data.	126
Data table features useful	Hiding, restoring, ordering and colouring the columns and values in the data table improves the usability of the table.	72
Event concept and approach not clear	The purpose to gather data for particular events complicates the understanding of the event concept and terminology.	33
Export (meta)data useful	Exporting the data, once it is understood, is useful for a subsequent analysis and the metadata for the provenance and reusability.	131
Home tab elements useful	The text, diagram and links in the home tab facilitate the overall understanding of processes underpinned by the UI.	50

Metadata content clear and useful	The content from the data provenance, data lineage and/or full metadata windows is clear and/or useful.	88
Metadata content not useful or confusing	The information provided in the data provenance table, data lineage and/or full metadata exploration windows is not useful or confusing.	71
Moderator assist	The moderator intervenes because the participant struggles to complete a task, wanders off task or goes too deep into a task or the system crashes.	174
Navigation complex	The design of the export and metadata buttons together with the multiple tab approach, the functionality of the pop ups and some of the data table features complicate the navigation.	151
Overall positive experience	Generic comment on the positive experience when using the tool.	48
Plot design complex	The content, axis ticks and labels of the plots are not clear.	60
Query inputs and elements useful	The query input options are clear in general and the tool tips, text and drop downs (multi)inputs help the user in understanding the query process.	191
Query process clear	The query process and execution is clear.	62
Query process complex	The query process and the sequence and meaning of the query inputs are not understood.	50
Search feature simple	Control+F function to look for specific elements in the metadata is simple and easy for the user.	61
Search information process complex	Finding relevant information in the metadata windows can be complex.	32

Simple export	Data and metadata generated are simple and easy to export.	70
System issue	The user identifies a system issue as taking long to load, crash or not responsive.	45
Task sequence complex	The task was confusing, not completed in the required order or had to be read again.	156
Visualisation useful	Visualising the data as a heat map, box plot and/or polar plot is useful, simple, interactivity adds value and the plots can be compared.	125
Wording not clear	A word or expression that appears in a paragraph is not clear or used appropriately.	75

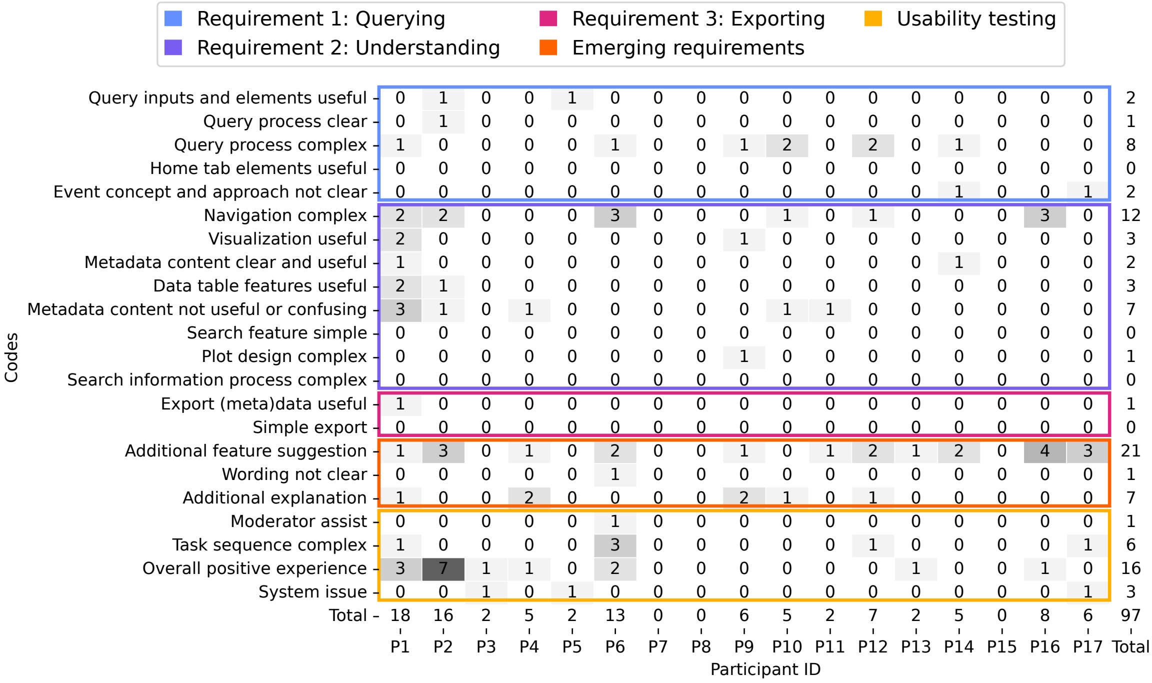
Table D.3: Description of the codes and references from the participants in the thematic analysis of the usability sessions transcripts (P3).

Code	**Description**	**References**
Complex report content	The amount of text in the interface and report can be reduced and the metadata section highlighted	26
CSV output with missing data	Some of the CSV files have missing data and the reason is not clear	26
Data availability comment	Comment related to the type, resolution, granularity or coverage of the available environmental data.	27
Data linkage process clear	The data linkage process is clear from uploading the health data to exporting the linked data	229
Data protection importance	Data protection is important aspect when linking health data with other sources and for publication	20
Experimental methodology	The participant asks if the task is completed or goes back to read the task again and comments on the lack of preparation for the session	35

Improved tool usability and output data	The tool is easy to use, time saving and better than before. The output data is in a usable format to conduct research.	155
Improvement to text or feature	Clarify wording, fix typo on a word, add extra information (e.g. tooltip), explain z-score interpretation bias and improve linking time progress	111
Moderator guidance	The moderator intervenes because the participant asks a direct question, wanders off task or the system has a technical issue.	125
Output data ready for analysis	No additional information is required to start the data analysis of the linked data as the dataset structure and content are clear and complete	99
Output data unclear	The content or structure of the output datasets and the filename are not enough to understand the meaning of the data	72
Positive comments on text	The text provided in the web interface is useful, helpful and consistent	32
Real use case applicable	The tool and data can be applied and useful for a real use case	54
Report helpful	The report is nice and useful with helpful features to explore and understand the linked data	169
Researcher's expertise	The researcher comments on their fit or background to evaluate the tool.	25
Some important features unclear	Uploading the (meta)data is not clear, done twice or not expected; and the linkage options are selected partially, not in order or without valid inputs.	144
Suggestions for additional features	Check upload input. Filter data table by variable. Add sanity checks. Select specific variables before linkage. Include feedback to improve data coverage. Allow for multiple aggregation methods. Offline deploy with own environmental data. Add research publications.	63

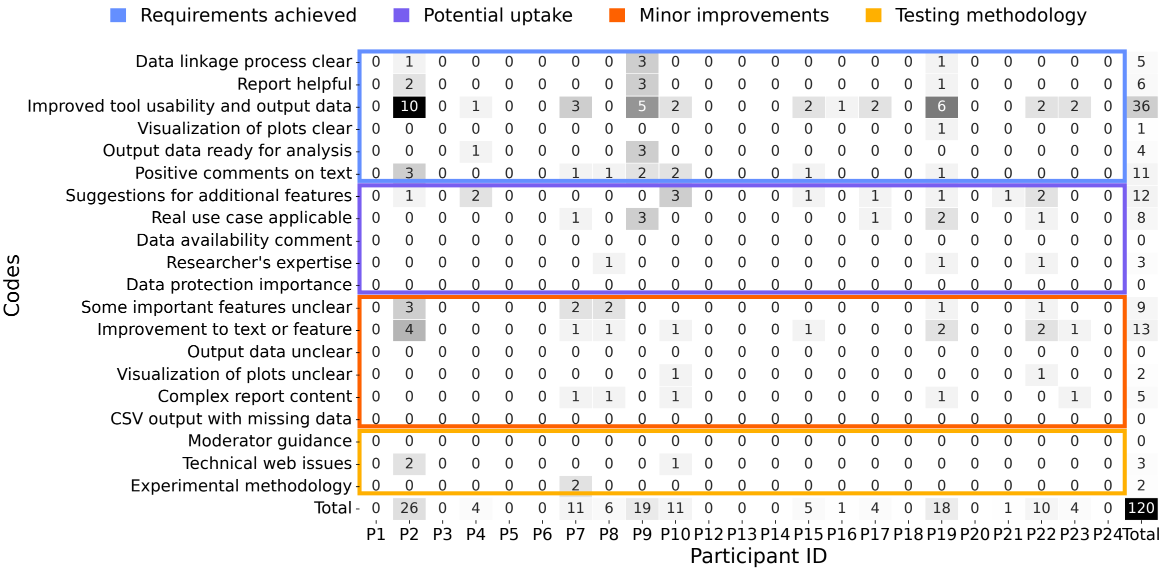
Technical web issues	Technical issues when editing cells in data tables, due to browser settings or the recording platform. The web needs to be refreshed to export the output or continue the session.	42
Visualisation of plots clear	The presentation of the variables and functionalities are clear in the plot	112
Visualisation of plots unclear	A component of the plot is not clear	50

E PSSUQ open comments thematic analysis



Codes	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	Total
Query inputs and elements useful	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2
Query process clear	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Query process complex	1	0	0	0	0	1	0	0	1	2	0	2	0	1	0	0	0	8
Home tab elements useful	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Event concept and approach not clear	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2
Navigation complex	2	2	0	0	0	3	0	0	0	1	0	1	0	0	0	3	0	12
Visualization useful	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3
Metadata content clear and useful	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2
Data table features useful	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Metadata content not useful or confusing	3	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	7
Search feature simple	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Plot design complex	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Search information process complex	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Export (meta)data useful	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Simple export	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Additional feature suggestion	1	3	0	1	0	2	0	0	1	0	1	2	1	2	0	4	3	21
Wording not clear	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Additional explanation	1	0	0	2	0	0	0	0	2	1	0	1	0	0	0	0	0	7
Moderator assist	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Task sequence complex	1	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0	1	6
Overall positive experience	3	7	1	1	0	2	0	0	0	0	0	0	1	0	0	1	0	16
System issue	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	3
Total	18	16	2	5	2	13	0	0	6	5	2	7	2	5	0	8	6	97

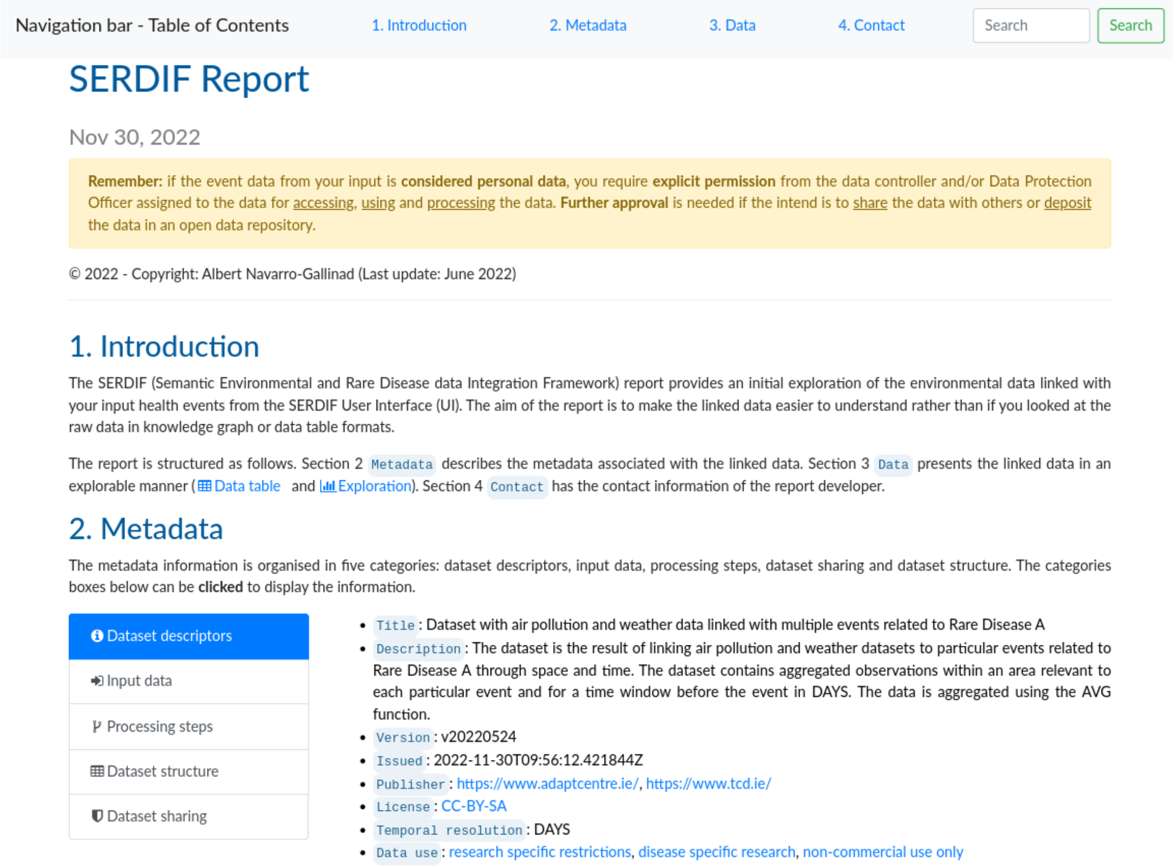
Figure E.1: Categorization of the PSSUQ open comments using the same codes as in the usability sessions transcripts displaying coherence between both sources (P2).



Codes	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	Total
Data linkage process clear	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	5
Report helpful	0	2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	6
Improved tool usability and output data	0	10	0	1	0	0	3	0	5	2	0	0	0	2	1	2	0	6	0	0	2	2	0	36
Visualization of plots clear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Output data ready for analysis	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Positive comments on text	0	3	0	0	0	0	1	1	2	2	0	0	0	1	0	0	0	1	0	0	0	0	0	11
Suggestions for additional features	0	1	0	2	0	0	0	0	0	3	0	0	0	1	0	1	0	1	0	1	2	0	0	12
Real use case applicable	0	0	0	0	0	0	1	0	3	0	0	0	0	0	0	1	0	2	0	0	1	0	0	8
Data availability comment	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Researcher's expertise	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	3
Data protection importance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Some important features unclear	0	3	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	9
Improvement to text or feature	0	4	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	2	0	0	2	1	0	13
Output data unclear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Visualization of plots unclear	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	2
Complex report content	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	5
CSV output with missing data	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moderator guidance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Technical web issues	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Experimental methodology	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Total	0	26	0	4	0	0	11	6	19	11	0	0	0	5	1	4	0	18	0	1	10	4	0	120

Figure E.2: Categorization of the PSSUQ open comments using the same codes as in the usability sessions transcripts displaying coherence between both sources (P3).

F SERDIF interactive report



SERDIF Report

Nov 30, 2022

Remember: if the event data from your input is **considered personal data**, you require **explicit permission** from the data controller and/or Data Protection Officer assigned to the data for accessing, using and processing the data. **Further approval** is needed if the intend is to share the data with others or deposit the data in an open data repository.

© 2022 - Copyright: Albert Navarro-Gallinad (Last update: June 2022)

1. Introduction

The SERDIF (Semantic Environmental and Rare Disease data Integration Framework) report provides an initial exploration of the environmental data linked with your input health events from the SERDIF User Interface (UI). The aim of the report is to make the linked data easier to understand rather than if you looked at the raw data in knowledge graph or data table formats.

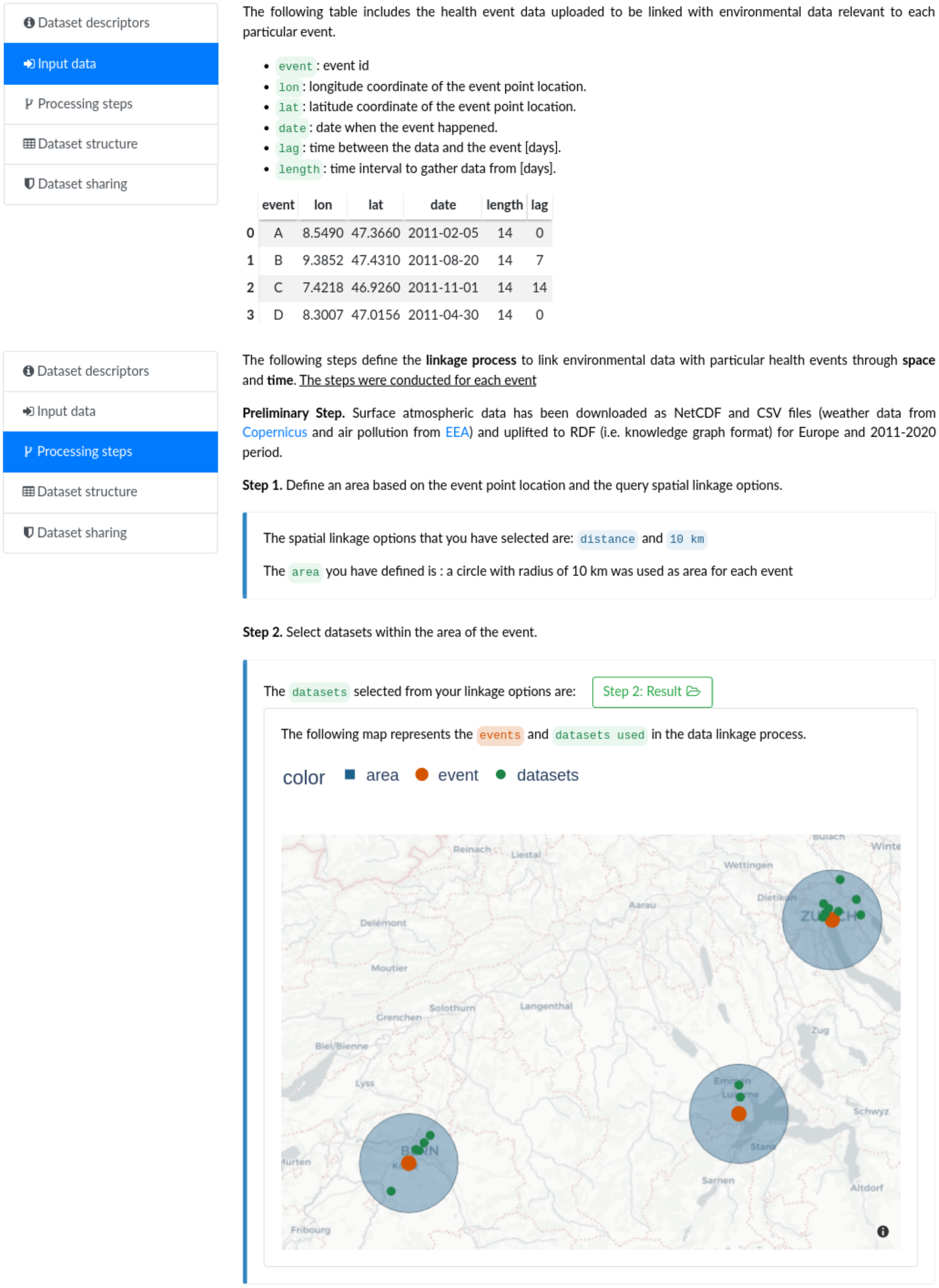
The report is structured as follows. Section 2 `Metadata` describes the metadata associated with the linked data. Section 3 `Data` presents the linked data in an explorable manner (Data table and Exploration). Section 4 `Contact` has the contact information of the report developer.

2. Metadata

The metadata information is organised in five categories: dataset descriptors, input data, processing steps, dataset sharing and dataset structure. The categories boxes below can be **clicked** to display the information.

- `Title` : Dataset with air pollution and weather data linked with multiple events related to Rare Disease A
- `Description` : The dataset is the result of linking air pollution and weather datasets to particular events related to Rare Disease A through space and time. The dataset contains aggregated observations within an area relevant to each particular event and for a time window before the event in DAYS. The data is aggregated using the AVG function.
- `Version` : v20220524
- `Issued` : 2022-11-30T09:56:12.421844Z
- `Publisher` : https://www.adaptcentre.ie/, https://www.tcd.ie/
- `License` : CC-BY-SA
- `Temporal resolution` : DAYS
- `Data use` : research specific restrictions, disease specific research, non-commercial use only

Figure F.3: Screenshot of the interactive report generated as a result of using the SERDIF user interface - 1.



The following table includes the health event data uploaded to be linked with environmental data relevant to each particular event.

- `event` : event id
- `lon` : longitude coordinate of the event point location.
- `lat` : latitude coordinate of the event point location.
- `date` : date when the event happened.
- `lag` : time between the data and the event [days].
- `length` : time interval to gather data from [days].

	event	lon	lat	date	length	lag
0	A	8.5490	47.3660	2011-02-05	14	0
1	B	9.3852	47.4310	2011-08-20	14	7
2	C	7.4218	46.9260	2011-11-01	14	14
3	D	8.3007	47.0156	2011-04-30	14	0

The following steps define the **linkage process** to link environmental data with particular health events through **space** and **time**. The steps were conducted for each event

Preliminary Step. Surface atmospheric data has been downloaded as NetCDF and CSV files (weather data from Copernicus and air pollution from EEA) and uplifted to RDF (i.e. knowledge graph format) for Europe and 2011-2020 period.

Step 1. Define an area based on the event point location and the query spatial linkage options.

The spatial linkage options that you have selected are: `distance` and `10 km`

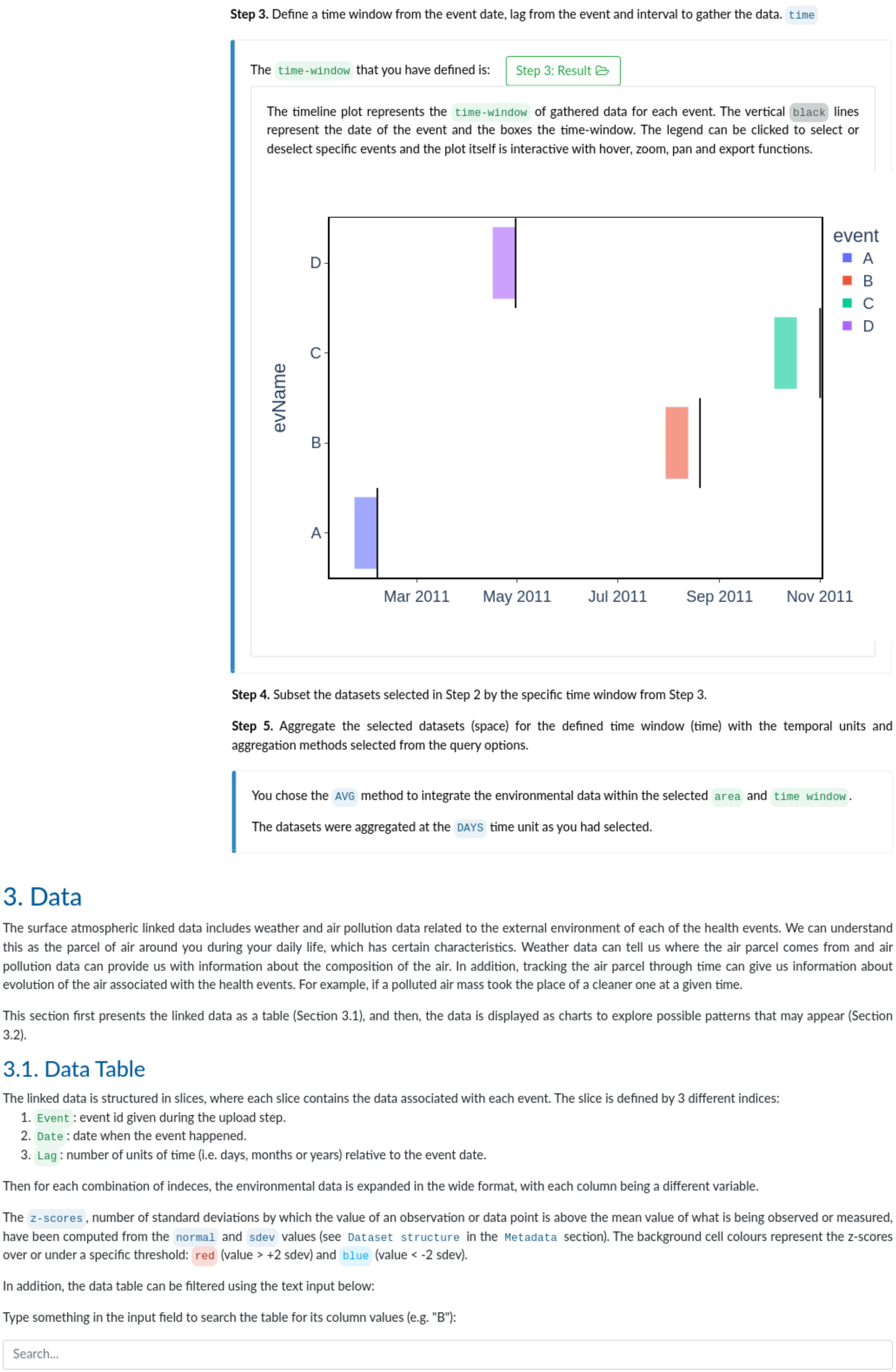
The `area` you have defined is : a circle with radius of 10 km was used as area for each event

Step 2. Select datasets within the area of the event.

The `datasets` selected from your linkage options are:

The following map represents the `events` and `datasets used` in the data linkage process.

Figure F.4: Screenshot of the interactive report generated as a result of using the SERDIF user interface - 2.



Step 3. Define a time window from the event date, lag from the event and interval to gather the data. `time`

The `time-window` that you have defined is: Step 3: Result

The timeline plot represents the `time-window` of gathered data for each event. The vertical `black` lines represent the date of the event and the boxes the time-window. The legend can be clicked to select or deselect specific events and the plot itself is interactive with hover, zoom, pan and export functions.

Step 4. Subset the datasets selected in Step 2 by the specific time window from Step 3.

Step 5. Aggregate the selected datasets (space) for the defined time window (time) with the temporal units and aggregation methods selected from the query options.

You chose the `AVG` method to integrate the environmental data within the selected `area` and `time window`.

The datasets were aggregated at the `DAYS` time unit as you had selected.

3. Data

The surface atmospheric linked data includes weather and air pollution data related to the external environment of each of the health events. We can understand this as the parcel of air around you during your daily life, which has certain characteristics. Weather data can tell us where the air parcel comes from and air pollution data can provide us with information about the composition of the air. In addition, tracking the air parcel through time can give us information about evolution of the air associated with the health events. For example, if a polluted air mass took the place of a cleaner one at a given time.

This section first presents the linked data as a table (Section 3.1), and then, the data is displayed as charts to explore possible patterns that may appear (Section 3.2).

3.1. Data Table

The linked data is structured in slices, where each slice contains the data associated with each event. The slice is defined by 3 different indices:

1. `Event` : event id given during the upload step.
2. `Date` : date when the event happened.
3. `Lag` : number of units of time (i.e. days, months or years) relative to the event date.

Then for each combination of indeces, the environmental data is expanded in the wide format, with each column being a different variable.

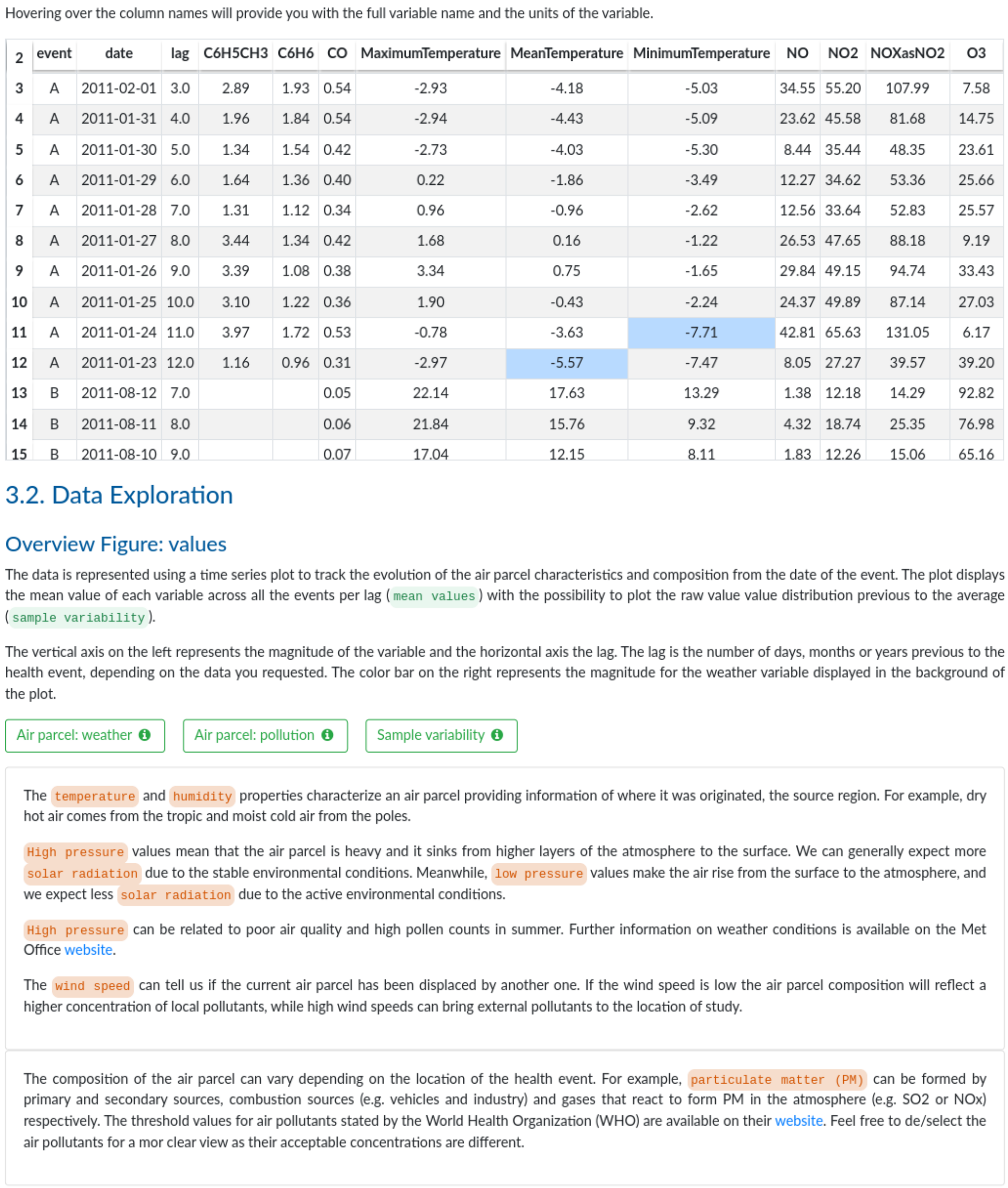
The `z-scores`, number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured, have been computed from the `normal` and `sdev` values (see `Dataset structure` in the `Metadata` section). The background cell colours represent the z-scores over or under a specific threshold: `red` (value > +2 sdev) and `blue` (value < -2 sdev).

In addition, the data table can be filtered using the text input below:

Type something in the input field to search the table for its column values (e.g. "B"):

Search...

Figure F.5: Screenshot of the interactive report generated as a result of using the SERDIF user interface - 3.



Hovering over the column names will provide you with the full variable name and the units of the variable.

2	event	date	lag	C6H5CH3	C6H6	CO	MaximumTemperature	MeanTemperature	MinimumTemperature	NO	NO2	NOXasNO2	O3
3	A	2011-02-01	3.0	2.89	1.93	0.54	-2.93	-4.18	-5.03	34.55	55.20	107.99	7.58
4	A	2011-01-31	4.0	1.96	1.84	0.54	-2.94	-4.43	-5.09	23.62	45.58	81.68	14.75
5	A	2011-01-30	5.0	1.34	1.54	0.42	-2.73	-4.03	-5.30	8.44	35.44	48.35	23.61
6	A	2011-01-29	6.0	1.64	1.36	0.40	0.22	-1.86	-3.49	12.27	34.62	53.36	25.66
7	A	2011-01-28	7.0	1.31	1.12	0.34	0.96	-0.96	-2.62	12.56	33.64	52.83	25.57
8	A	2011-01-27	8.0	3.44	1.34	0.42	1.68	0.16	-1.22	26.53	47.65	88.18	9.19
9	A	2011-01-26	9.0	3.39	1.08	0.38	3.34	0.75	-1.65	29.84	49.15	94.74	33.43
10	A	2011-01-25	10.0	3.10	1.22	0.36	1.90	-0.43	-2.24	24.37	49.89	87.14	27.03
11	A	2011-01-24	11.0	3.97	1.72	0.53	-0.78	-3.63	-7.71	42.81	65.63	131.05	6.17
12	A	2011-01-23	12.0	1.16	0.96	0.31	-2.97	-5.57	-7.47	8.05	27.27	39.57	39.20
13	B	2011-08-12	7.0			0.05	22.14	17.63	13.29	1.38	12.18	14.29	92.82
14	B	2011-08-11	8.0			0.06	21.84	15.76	9.32	4.32	18.74	25.35	76.98
15	B	2011-08-10	9.0			0.07	17.04	12.15	8.11	1.83	12.26	15.06	65.16

3.2. Data Exploration

Overview Figure: values

The data is represented using a time series plot to track the evolution of the air parcel characteristics and composition from the date of the event. The plot displays the mean value of each variable across all the events per lag (`mean values`) with the possibility to plot the raw value value distribution previous to the average (`sample variability`).

The vertical axis on the left represents the magnitude of the variable and the horizontal axis the lag. The lag is the number of days, months or years previous to the health event, depending on the data you requested. The color bar on the right represents the magnitude for the weather variable displayed in the background of the plot.

The `temperature` and `humidity` properties characterize an air parcel providing information of where it was originated, the source region. For example, dry hot air comes from the tropic and moist cold air from the poles.

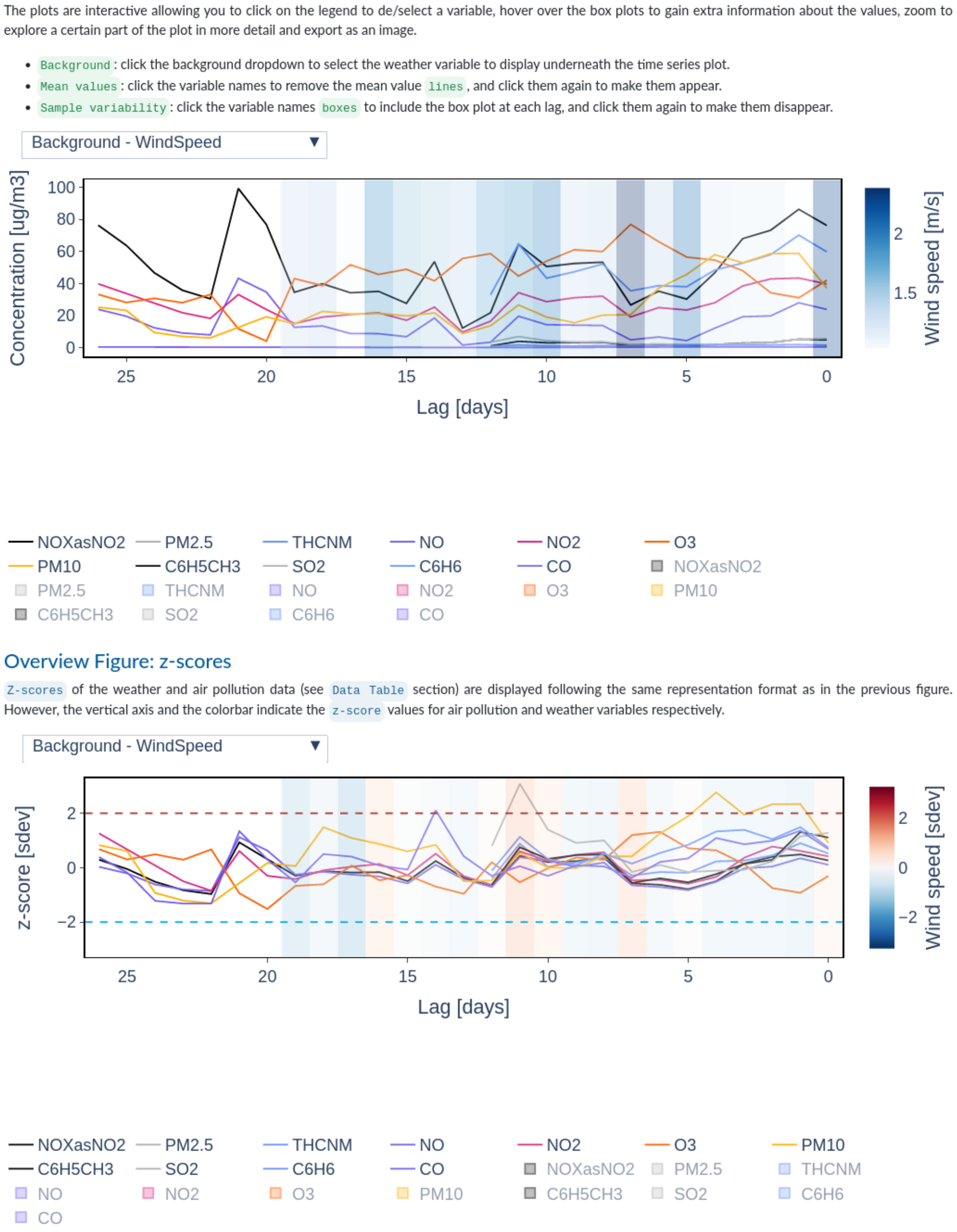
`High pressure` values mean that the air parcel is heavy and it sinks from higher layers of the atmosphere to the surface. We can generally expect more `solar radiation` due to the stable environmental conditions. Meanwhile, `low pressure` values make the air rise from the surface to the atmosphere, and we expect less `solar radiation` due to the active environmental conditions.

`High pressure` can be related to poor air quality and high pollen counts in summer. Further information on weather conditions is available on the Met Office website.

The `wind speed` can tell us if the current air parcel has been displaced by another one. If the wind speed is low the air parcel composition will reflect a higher concentration of local pollutants, while high wind speeds can bring external pollutants to the location of study.

The composition of the air parcel can vary depending on the location of the health event. For example, `particulate matter (PM)` can be formed by primary and secondary sources, combustion sources (e.g. vehicles and industry) and gases that react to form PM in the atmosphere (e.g. SO2 or NOx) respectively. The threshold values for air pollutants stated by the World Health Organization (WHO) are available on their website. Feel free to de/select the air pollutants for a mor clear view as their acceptable concentrations are different.

Figure F.6: Screenshot of the interactive report generated as a result of using the SERDIF user interface - 4.



The plots are interactive allowing you to click on the legend to de/select a variable, hover over the box plots to gain extra information about the values, zoom to explore a certain part of the plot in more detail and export as an image.

- `Background` : click the background dropdown to select the weather variable to display underneath the time series plot.
- `Mean values` : click the variable names to remove the mean value `lines` , and click them again to make them appear.
- `Sample variability` : click the variable names `boxes` to include the box plot at each lag, and click them again to make them disappear.

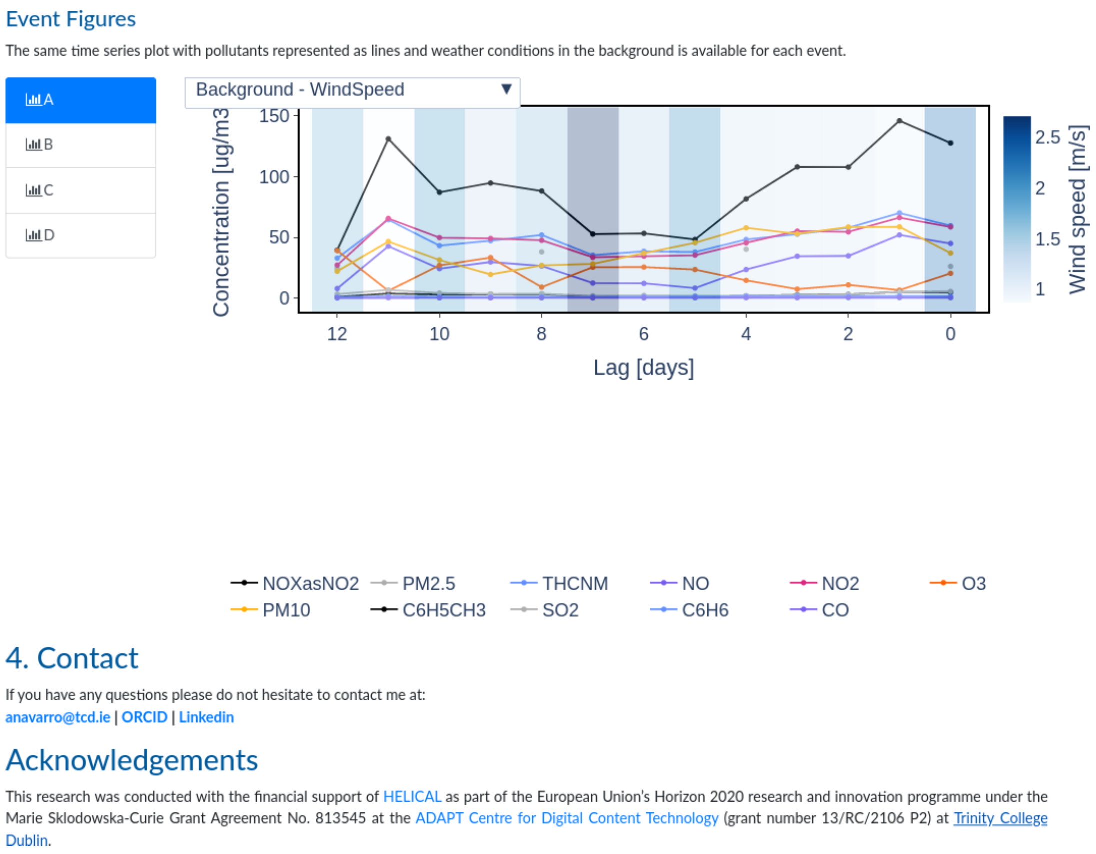
Background - WindSpeed ▼

Overview Figure: z-scores

`Z-scores` of the weather and air pollution data (see `Data Table` section) are displayed following the same representation format as in the previous figure. However, the vertical axis and the colorbar indicate the `z-score` values for air pollution and weather variables respectively.

Background - WindSpeed ▼

Figure F.7: Screenshot of the interactive report generated as a result of using the SERDIF user interface - 5.



Event Figures

The same time series plot with pollutants represented as lines and weather conditions in the background is available for each event.

4. Contact

If you have any questions please do not hesitate to contact me at:
anavarro@tcd.ie | ORCID | Linkedin

Acknowledgements

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 Sklodowska-Curie Grant Agreement No. 813545 at the ADAPT Centre for Digital Content Technology (grant number 13/RC/2106 P2) at Trinity College Dublin.

Figure F.8: Screenshot of the interactive report generated as a result of using the SERDIF user interface - 5.