Engaging Content
Engaging People

A World Leading SFI Research Centre

Science Foundation Ireland For what's next

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

Authors: Albert Navarro-Gallinad, Fabrizio Orlandi and Declan O’Sullivan
ADAPT Centre for Digital Content, Trinity College Dublin, Dublin, Ireland

International Joint Conference on Knowledge Graphs – 08-12-2021

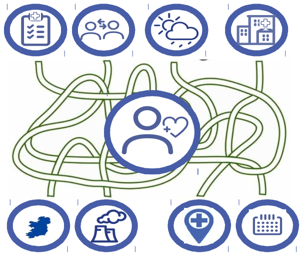
HOST INSTITUTION
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

HOST INSTITUTION
DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

PARTNER INSTITUTIONS
TU Dublin
UCD Dublin
CIT Cork Institute of Technology Institiúid Teicneolaíochta Chorcaí
AIT Connect & Discover
Maynooth University National University of Ireland Maynooth

FUNDED BY
Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union European Regional Development Fund

Research Background – Problems Addressed and Impact



Interoperability Challenge

Linking health data with scientific data through location and time

Use Of Knowledge Graphs

Non-technical researchers using and navigating Knowledge Graphs to answer complex research questions

Better Quality of Life

Researchers will be able to understand better diseases, which could lead to new or better treatments

State Of the Art Review

Review on: combining methods for rare disease clinical data with other data sources using Semantic Web technologies.

State Of the Art

Biomedical domain

Linking data from:

- Biobanks and registries
- Genetic and epigenetic data

Visual interfaces

Some required Semantic Web practical expertise (e.g. SPARQL queries) !

Usability studies

Only one usability study using a customized approach !

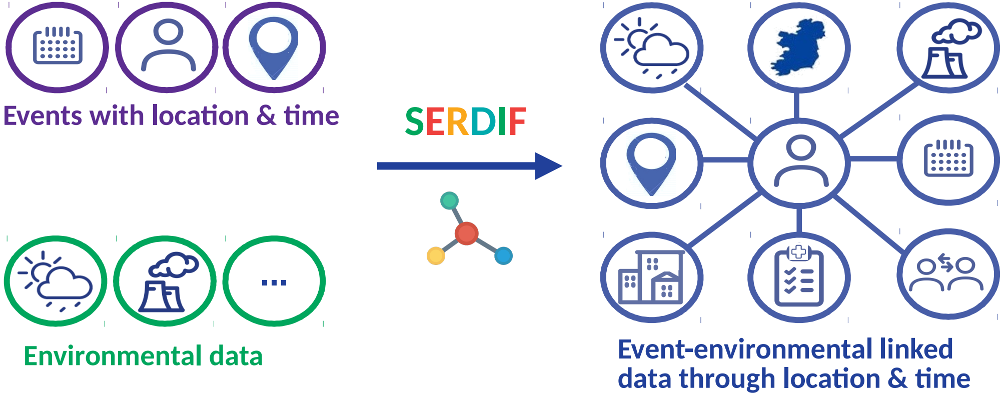
✗ Limited studies combining health and environmental data together using KGs.

✗ Limited usability studies with standard metrics in the evaluation.

What did we do? – Design the SERDIF approach

SERDIF (Semantic Environmental and Rare disease Data Integration Framework)

- Informed from the SOA results and initial Health Data Researchers (HDR) requirements*


Events with location & time
Environmental data
...
SERDIF
Event-environmental linked data through location & time

The framework is a combination of (1) a methodology, (2) a knowledge graph and (3) a dashboard.

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

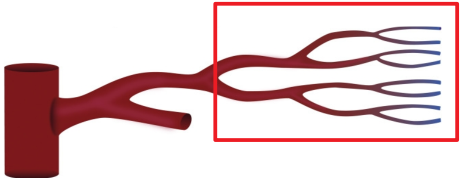
Research Implementation – Methodology & Use Case

Methodology:

1. Data Collection
2. Semantic Uplift
3. Data Query and Filter
4. Data Visualization
5. Data Export/Downlift
6. Usability Evaluation

AAV* in Ireland

ANCA-Associated vasculitis (AAV) is a rare autoimmune diseased of unknown aetiology.



Genetically susceptible

Epigenetic factors

Environmental factors

...

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

Research Implementation – Data Query & Filter

Methodology:

1 Data Collection

2 Semantic Uplift

3 Data Query and Filter

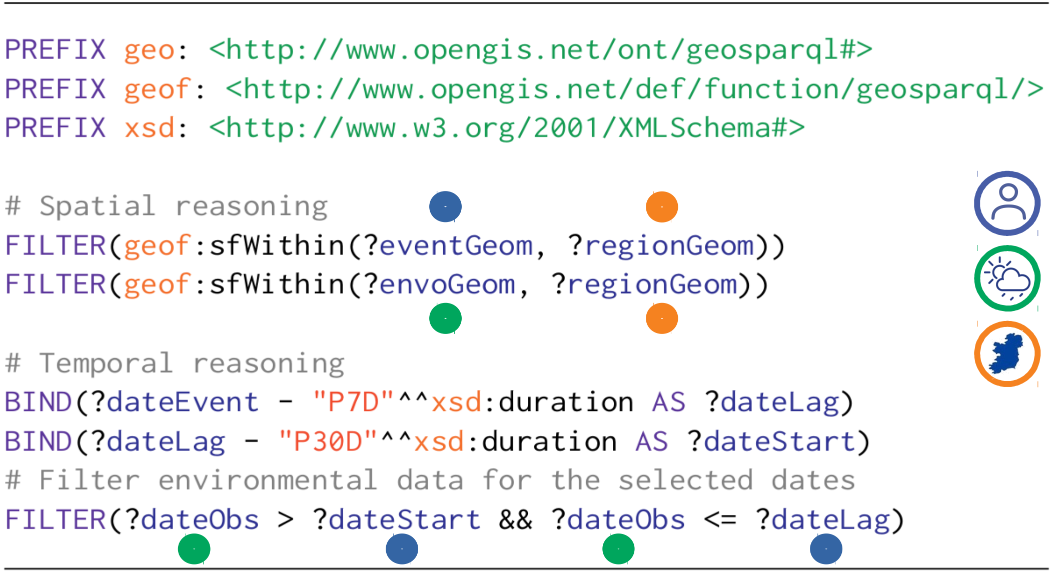
4 Data Visualization

5 Data Export/Downlift

6 Usability Evaluation

Options for linking different types of RDF graphs:

3. Spatial and temporal reasoning at the SPARQL query level



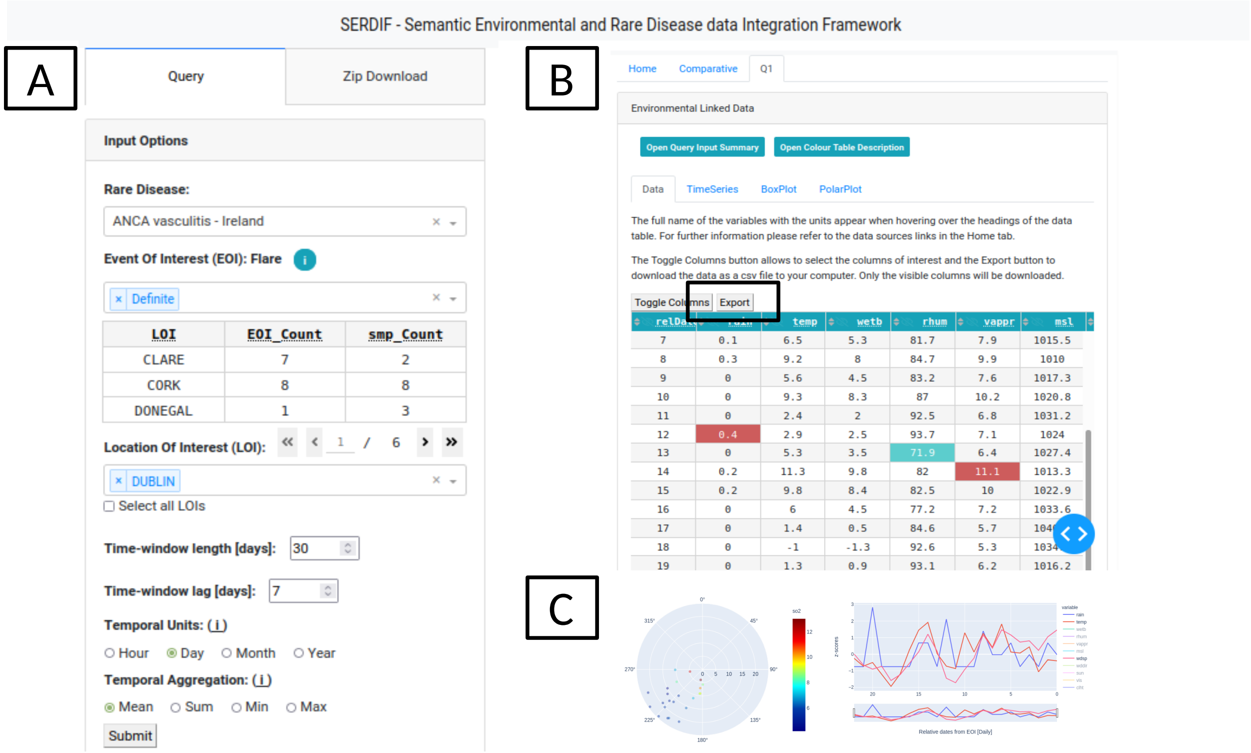
```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

# 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)
```

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

Research Implementation – Data Visualization

Methodology:

1. Data Collection
2. Semantic Uplift
3. Data Query and Filter
4. Data Visualization
5. Data Export/Downlift
6. Usability Evaluation



SERDIF - Semantic Environmental and Rare Disease data Integration Framework

A

Query | Zip Download

Input Options

Rare Disease: ANCA vasculitis - Ireland

Event Of Interest (EOI): Flare

Definite

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

1 / 6

Location Of Interest (LOI): DUBLIN

□ Select all LOIs

Time-window length [days]: 30

Time-window lag [days]: 7

Temporal Units: (i)

○ Hour ● Day ○ Month ○ Year

Temporal Aggregation: (i)

● Mean ○ Sum ○ Min ○ Max

Submit

B

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

relDa	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	104
18	0	-1	-1.3	92.6	5.3	1034
19	0	1.3	0.9	93.1	6.2	1016.2

C

Github: https://github.com/navarral/ijckg2021-serdif-paper | **Dashboard:** https://serdif.adaptcentre.ie/dashboard

Research Implementation – Experiment Setup

Methodology:

1 Data Collection

2 Semantic Uplift

3 Data Query and Filter

4 Data Visualization

5 Data Export/Downlift

6 Usability Evaluation

Health Data Researchers (HDR)

General aspects

- □ Sample size 10
- □ Videoconference
- □ Tasks derived from HDR consensus
- □ Think aloud protocol

Quantitative metrics

- □ Task completion
- □ Time per task
- □ PSSUQ

Qualitative metrics

- □ Session transcript
- □ Thematic analysis

Research Implementation – Usability Evaluation

Methodology:

1. Data Collection
2. Semantic Uplift
3. Data Query and Filter
4. Data Visualization
5. Data Export/Downlift
6. Usability Evaluation

AAV in Ireland first iteration of the evaluation:

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

Lessons Learnt

Health data domain

- Gaining access to health data can be complicated and long
- Bidirectional communication with domain experts is key
- Jargon as simple as possible

Technical aspects

- Reusing vocabularies/ ontologies facilitates the semantic uplift and data reuse
- Queries execution time can be improved if the query is broken down into smaller pieces

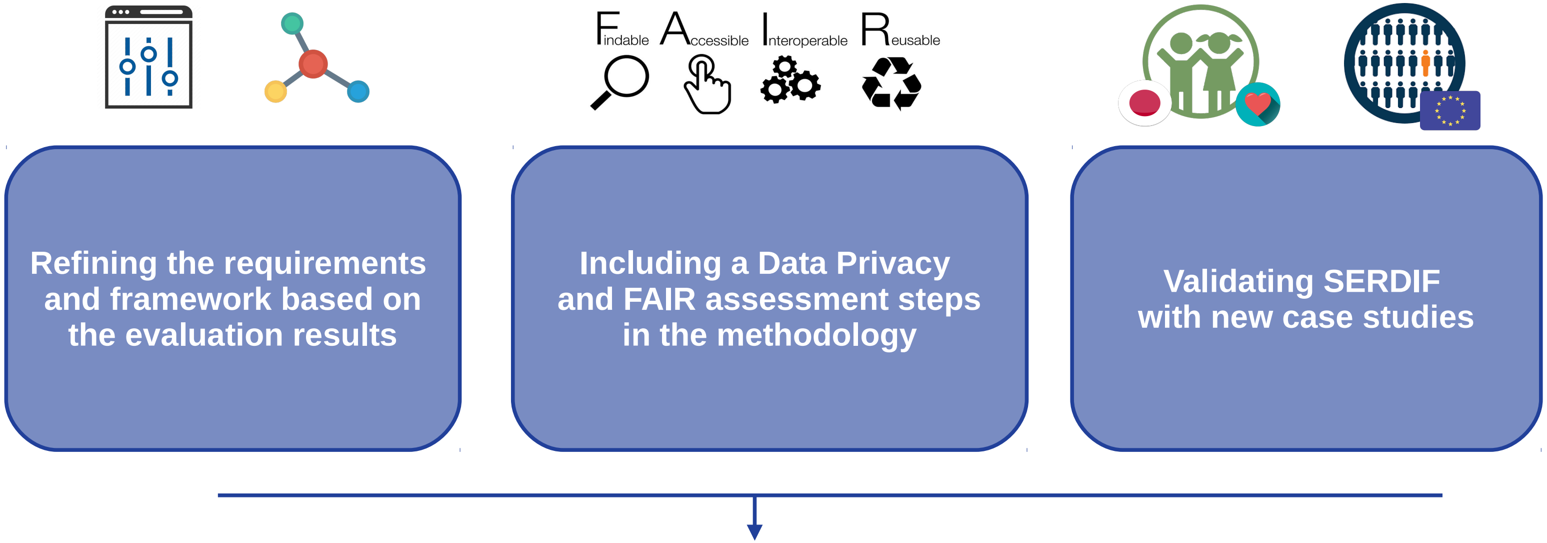
Usability testing

- Double check if the participants have read the documentation
- Videoconferencing environment facilitated the think aloud protocol
- Thematic analysis and PSSUQ are valuable metrics

General aspects

- If researchers find the tool useful for them, they tend to give more relevant feedback
- Examples are great to get the message across

Current Work



Refining the requirements and framework based on the evaluation results

Including a Data Privacy and FAIR assessment steps in the methodology

Validating SERDIF with new case studies

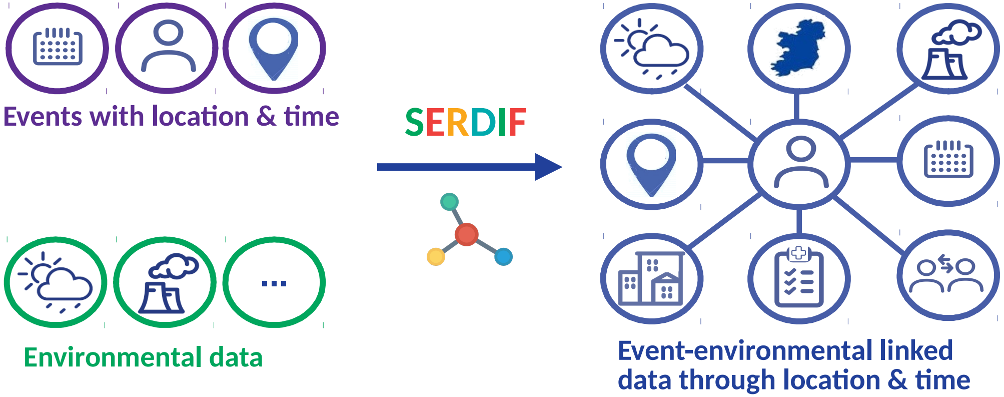
- Substituting the SOSA vocabulary for RDF data cube
- Adding steps to the methodology: Data Privacy and FAIR assessment (First Draft under preparation)
- Undertaking further validation of SERDIF with: Kawasaki Disease in Japan and Vasculitis disease in Europe

Paper Contributions

1 SERDIF

2 Linking disciplines

3 In-use application



Events with location & time
Environmental data
...
SERDIF
Event-environmental linked data through location & time

Preprint available: http://hdl.handle.net/2262/97660 | **Slides:** https://github.com/navarral/ijckg2021-serdif-paper

Albert Navarro Gallinad
albert.navarro@adaptcentre.ie
PhD Student

https://www.adaptcentre.ie/
http://helical-itn.eu/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813545 and No. 713567 at the ADAPT Centre for Digital Content Technology (grant number 13/RC/2106_P2) at Trinity College Dublin.