Privacy-Preserving Dashboard for F.A.I.R Head and Neck Cancer data supporting multi-centered collaborations

Varsha Gouthamchand[1,2,*], Ananya Choudhury[1,2], Frank Hoebers[1], Frederik Wesseling[1], Mattea Welch[3], Sejin Kim[3], Benjamin Haibe-Kains[3], Joanna Kazmierska[4], Andre Dekker[1,2], Johan van Soest[5,1] and Leonard Wee[1,2]

[1]*Dept of Radiation Oncology (Maastro), GROW School of Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands*

[2]*Clinical Data Science, Faculty of Health Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands*

[3]*Radiation Medicine Program, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada*

[4]*Dept of Radiation Oncology, Greater Poland Cancer Centre II, Poznan, Poland*

[5]*Brightlands Institute for Smart Society, Faculty of Science and Engineering, Maastricht University, Heerlen, The Netherlands*

Abstract

Research in modern healthcare requires vast volumes of data from various healthcare centers across the globe. It is not always feasible to centralize clinical data without compromising privacy. A tool addressing these issues and facilitating reuse of clinical data is the need of the hour. The Federated Learning approach, governed in a set of agreements such as the Personal Health Train (PHT) manages to tackle these concerns by distributing models to the data centers instead of the traditional approach of centralizing datasets. One of the prerequisites of PHT is using semantically interoperable datasets for the models to be able to find them. FAIR (Findable, Accessible, Interoperable, Reusable) principles help in building interoperable and reusable data by adding knowledge representation and providing descriptive metadata. However, the process of making data FAIR is not always easy and straight-forward. Our main objective is to disentangle this process by using domain and technical expertise and get data prepared for federated learning. This paper introduces applications that are easily deployable as Docker containers, which will automate parts of the aforementioned process and significantly simplify the task of creating FAIR clinical data. Our method bypasses the need for clinical researchers to have a high degree of technical skills. We demonstrate the FAIR-ification process by applying it to five Head and Neck cancer datasets (four public and one private). The PHT paradigm is explored by building a distributed visualization dashboard from the aggregated summaries of the FAIR-ified datasets. Using the PHT infrastructure for exchanging only statistical summaries or model coefficients allows researchers to explore data from multiple centers without breaching privacy.

Keywords

FAIR, Knowledge graphs, Linked Data, Semantic Web, Ontologies, SPARQL, RDF, Federated Learning,

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences

[*]Corresponding author.

varsha.gouthamchand@maastro.nl (V. Gouthamchand)

0000-0002-4756-2866 (V. Gouthamchand); 0000-0001-9847-8165 (A. Choudhury); 0000-0002-4317-9181 (F. Hoebers); 0000-0002-5887-9826 (M. Welch); 0000-0002-0422-7996 (A. Dekker); 0000-0003-2548-0330 (J. v. Soest); 0000-0003-1612-9055 (L. Wee)

1. Background

Real-world clinical data is defined as data relating to an individual person's health status and/or the delivery of healthcare to a population that is routinely collected from a variety of sources. This is increasingly being used as evidence of treatment effectiveness as well as to guide clinical decision-making through the development of predictive prognostic models [2].

Re-use of real-world clinical data at scale presents two challenges. First is a lack of syntactic interoperability, i.e., technical differences due to database organization and divergence of human languages. A more flexible alternative is an "open world" approach focusing on semantic interoperability where the data can be queried and retrieved by independent external researchers [13] without having to know details in advance, about database structure or native coding schema of the data. Second is that clinical data will be usually horizontally partitioned; healthcare institutions each own similar sets of data fields but exclusively on their own human subjects. Due to the highly sensitive nature of patient medical data, great care must be taken if shared. Concerns over patient confidentiality and data controllership implies that it is not always an attractive option to aggregate all individual-level data into a few centralized repositories.

A federated learning paradigm attempts to address some of the privacy concerns. A privacy-by-design paradigm, e.g., Personal Health Train (PHT) [14], exchanges only aggregated statistical information. A necessary consequence of using the PHT is that the data must first be made semantically interoperable for algorithms to analyze the data remotely and autonomously.

The FAIR (Findable, Accessible, Interoperable, and Reusable) data principles were developed to maximize the value of digital assets, including real world clinical data, and it further emphasizes making data interoperable for machine processors and not only for humans [20]. It is important to emphasize that FAIR data does not imply open data, and open data itself may not be FAIR. The purpose of FAIR is that a given community, e.g., researchers and cancer clinicians, can achieve a high degree of interoperability and reusability on each other's data [8].

The Linked Data [3] concept elegantly captures some of the needs of FAIR by assigning machine-readable unique resource identifiers (URIs) to data elements, as well as capturing the relationships between them. This attribute of linked data can be exploited to integrate disparate pools of data, even if they comprise different domains, e.g., clinical examinations and image-based biomarkers extracted from radiology scans. Independent but linked databases readable by machines over Hypertext Transfer Protocol (HTTP) makes up a worldwide "Semantic Web" of FAIR data. Semantic web standards define a set of essential tools such as the Resource Descriptor Framework (RDF) [10] and a SPARQL Protocol and RDF Query Language (SPARQL) [12], for storing data and querying data, respectively. RDF represents data using a series of statements known as "triples", i.e., subject-predicate-object.

This work defines a partly automated procedure to make structured real-world patient data more FAIR according to Semantic Web standards. Specialist competencies are leveraged collaboratively at the right place, i.e., clinicians and data creators will focus on annotation using rich descriptions; domain experts and ontologists can encode knowledge representation with data graph construction, and data scientists can instantiate these annotations and integrate data from disparate sources using a singular SPARQL query. Privacy-preserving federated learning is used to generate a visualization of aggregated cohort statistics across five private and public datasets in head-and-neck cancer, without revealing individual-level data.

2. Implementation

A schematic illustration of the general workflow to convert structured raw data to FAIR semantic web data is provided in Fig 1. Our tooling consists of three principal parts, (1) a graphical user interface (GUI) to select data for serialization as RDF triples and for the data owner to attach some descriptive information, (2) a collaborative annotation step to attach one or more relevant domain ontologies and hence define a fit-for-purpose graph data structure, and (3) a means to query data across multiple FAIR data graphs via a single federated SPARQL query. We packaged these tools as Docker containers to make them platform-independent and easier to deploy; these are made open access (see Availability section).

2.1. Structured data conversion to FAIR data graph

The first step takes existing structured data and processes either comma-separated values (CSV) or relational databases (any generic SQL format) into RDF. This component is provided with a GUI running locally where (non-technical) data owners simply browse and choose their dataset for processing. Triplifier [11] is a Java-based resource (integrated with the GUI) that automatically serializes a table into RDF as a Terse-Triple Language (TTL) file [16] and compiles the schema of the ingested table as a database-specific Web Ontology Language (OWL) file [9]. The same GUI also requests data owners to add descriptions and metadata, such as data type of each field (continuous, discrete, categorical ordinal, or patient identifier). The data owner is also able to attach some preselected definitions to their data fields and/or include supplementary information in free text comments. The data owner's pre-annotations are directly written into the aforementioned OWL schema file. The resulting TTL and OWL files are automatically saved in a graph database (we included a free version of GraphDB in our Docker deployment).

2.2. Annotation of datasets with ontologies

The procedure was specifically designed to decouple the contents of the database (in the TTL) from the dictionary/coding of the database (in the OWL), thus allowing an external collaborator to work on semantic annotations using the OWL without needing to read the actual contents of the data, which does contain person-specific and highly privacy-sensitive information. The data owner may thus share the OWL file either publicly, or privately to a defined collaboration, of researchers, domain experts, clinicians and other data owners. For each distinct dataset, a semantically meaningful mapping of the database-specific entities onto a publicly accessible do-main ontology (e.g., ROO and NCIT) will be made through consensus and correspondence (e.g., through extended email discussions).

We provide Python scripts that add unique dataset specific annotations as a graph object ("annotation.local") directly into the local GraphDB. Importantly, if the use case changes (or if a new research question emerges) such that an alternative annotation is required, this can be easily implemented. Re-annotation of the data is always possible because it sits on top of the original OWL and TTL without needing to edit or modify any of the original schema and original contents. A SPARQL query external to the dataset then references the equivalencies in the "annotation.local" in order to extract the correct entities and relations specific to the dataset.

2.3. Tagged release of software

The code for this project is made open access on Github (refer Availability section); instructions for use have been provided in associated markdown documents. We have provided a Docker installation containing a Jupyter notebook, GraphDB and triplifier. All the python scripts needed are packaged into the distribution as Jupyter notebooks.

2.4. PHT infrastructure

An open source Vantage6 infrastructure [7] (v0.2.4) implementing the PHT method has been previously used to develop and validate a federated CPH model in anal cancer patients across three countries [15]. Full details of Vantage6 are given in its accompanying technical documentation [19].

2.5. Distributed dashboard aggregation of head and neck cancer data

A demonstration was previously made available as a preprint [4] and remains available from our GitHub repository as the "MedRxiv" branch. To recapitulate briefly, we had created an automated process where the clinical case-mix data in four different open access datasets on The Cancer Imaging Archive (TCIA) [1] were serialized using triplifier, and each set of TTL and OWL files were inserted into its respective GraphDB database. On top of each of these GraphDB databases, Python scripts were executed which inserted local annotations on top of the TTL and OWL files, utilizing class entities and predicates from the NCIT and ROO ontologies.

This work extends the previous by adding a hitherto unpublished private dataset (HN3). We used the aforementioned GUI and triplifier to serialize the data and create the custom semantic ontology annotation for its RDF graph. We placed each of these five datasets in geographically dispersed Ubuntu virtual machines, with unique public IP addresses and network firewalls, but all connected to a Vantage6 infra-structure (illustrated schematically in Fig 2). We distributed a single SPARQL query through the PHT infrastructure to obtain aggregated cohort statistics (e.g., mean, range, etc.) from the federated datasets, then presented these in two ways – (i) an interactive Python visual dashboard built from PlotLy and Dash libraries, and (ii) a case-mix summary data frame that could be downloaded from the Vantage6 aggregation server as a Comma-Separated Values (CSV) file.

3. Results

For open data, the interested reader can get the original clinical data frame directly from TCIA. An example fragment of TTL and OWL generated by triplifier is shown in Fig 3. Note that (for ease of understanding) we have compressed the namespaces using standard semantic web notation in the bottom-left corner of the figure.

A graphical representation of as-serialized TTL content for one subject is shown on the left side of Fig 4. For argument's sake, the original contents might not be syntactically usable to the reader (e.g., the labels are in the Dutch language) and are not yet semantically interoperable. On the right side of Fig 4, we show how dataset-specific annotations mapped to the ROO and NCIT ontologies (including new descriptive predicates) render this data more FAIR. The inserted annotations are strictly additive, i.e., it does not alter or over-write the as-serialized contents. For simplicity of visualization, we masked some of the schema classes and predicates auto generated by triplifier. One can readily look up the unique URIs - C25364, C28421 and C16576 in the NCIT and find definitions for “patient identifier”, “sex” and “female”, respectively. Likewise, with the ROO, the URIs P100061, P100018 and P100042 are resolved as predicates “has_identifier”, “has_biological_sex” and “has_value”, respectively. Where needed, numerical values may be supplemented by extra predicates and classes indicating the exact units of the measure, e.g., age in years, and follow-up time as intervals of days, or months, or years. Additionally, if concepts like dates (Date of Birth/Death) need to be re-formatted to an agreed style (e.g., DD/MM/YYYY), an appropriate formatting task can be sent using the Vantage6 server to the data nodes.

A snapshot of an interactive visual dashboard is given as Fig 5, and we retrieved a case-mix aggregated summary table directly from the Vantage6 server as a CSV file. The latter was then reformatted and tidied to provide Table 1 in Supplementary material.

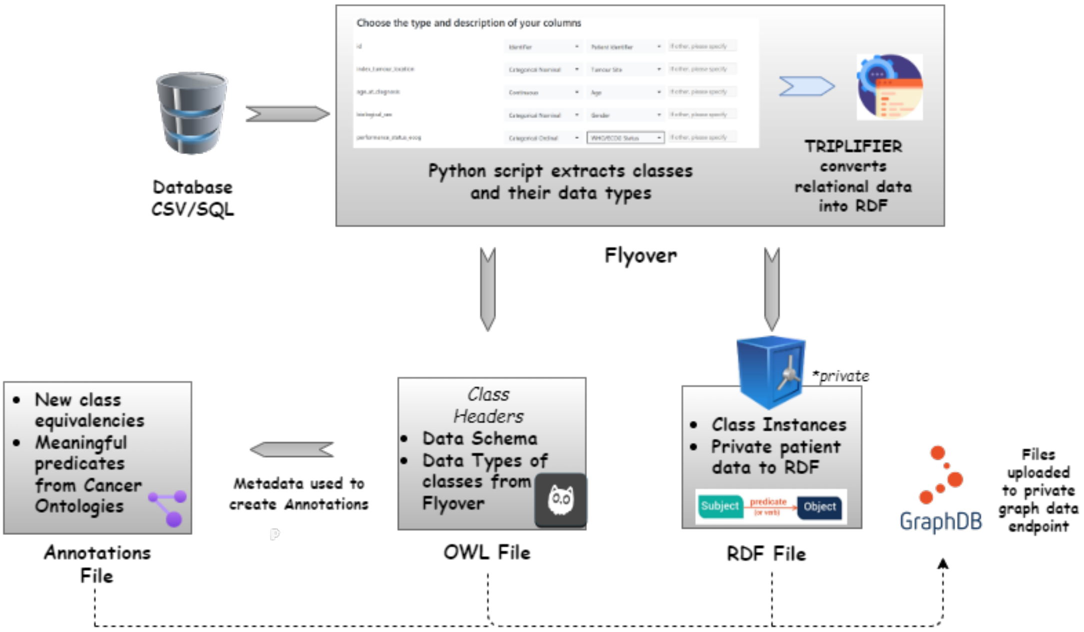
4. Discussion and Conclusion

We have produced a partly automated procedure that makes structured clinical data FAIR and available for distributed applications. Semantic interoperability and data linking has been achieved by local annotation with semantic ontologies, such that a single global SPARQL query using those ontologies will correctly filter the data.

This illustrative use case was selected to address how common clinically relevant questions may be addressed via privacy-preserving federated learning. In our distributed dashboard approach, we envisage that partners in an established collaboration will be able to safely explore and inter-compare each other’s private data repositories for suitable subsets of patients, without violating patient privacy. If open datasets are also annotated and made FAIR in the abovementioned manner, then published via an accessible web address, they can also be efficiently queried en masse with a single query in the above manner.

One of the most important and effective ways to make data FAIR is to assign a globally unique and persistent identifier to both the data repository and its linked metadata. The four open datasets here are unambiguously referenced using a Digital Object Identifier (DOI). Though private dataset HN3 is not openly accessible, we do openly disseminate the readable description of the dataset plus the schema (OWL) and its semantic ontology annotations (“annotation.local”) in open Zenodo repository with its unique persistent DOI for the metadata.

Assuming that re-casting the data into a universal master schema is not already done, our method proposes a flexible and adaptable means of applying semantic interoperability by means of annotation with an open semantic ontologies. In the Linked Data paradigm, each data entity as well as its relationship to other data entities is traceably and collectively mapped to a unique and persistent identifier. Every instance of the same identifier must mean semantic equivalence, entirely irrespective of the human-readable label, which is generally in the data owner’s own language. The ontologies not only establish the terminology and definitions but also include some knowledge representation that allows the possibility to apply machine-assisted logical inferencing.



Choose the type and description of your columns

id	Identifier	Patient Identifier	If other, please specify
index_tumour_location	Categorical Nominal	Tumour Site	If other, please specify
age_at_diagnosis	Continuous	Age	If other, please specify
biological_sex	Categorical Nominal	Gender	If other, please specify
performance_status_ecog	Categorical Ordinal	WHO/ECOG Status	If other, please specify

Database CSV/SQL
Python script extracts classes and their data types
TRIPLIFIER converts relational data into RDF
Flyover
New class equivalencies
Meaningful predicates from Cancer Ontologies
Annotations File
Metadata used to create Annotations
Class Headers
Data Schema
Data Types of classes from Flyover
OWL File
*private
Class Instances
Private patient data to RDF
Subject
predicate (or verb)
Object
RDF File
GraphDB
Files uploaded to private graph data endpoint

Figure 1: Making clinical data accessible as a FAIR graph database object

Availability

Project name: Flyover (tagged release: v1.0, preprint demonstration project branch: MedRxiv)
Project home page: https://github.com/MaastrichtU-CDS/projects_flyover_project
Zenodo Repository DOI: https://doi.org/10.5281/zenodo.7190551

Four public datasets were obtained from TCIA. RADIOMICS-HN1 [18] comprises clinical data, volumetric CT and PET of 137 patients with laryngeal carcinoma and OPC treated by RT alone or currently with either cisplatin or cetuximab. HNSCC contains clinical data and contrast-enhanced CT scans of 627 oropharyngeal cancer (OPC) patients [5]. OPC-Radiomics has clinical data and CT scans of 606 OPC subjects, treated by either radiotherapy or chemoradiotherapy between 2005 and 2010 [6]. HEAD-NECK-PET-CT [17] com-prised 298 subjects with multiple subsites of HNC each with clinical descriptors, PET and planning CT, treated between April 2006 and November 2014. The HN3 dataset is not publicly available at the present time due to material that is potentially identifiable to an individual.

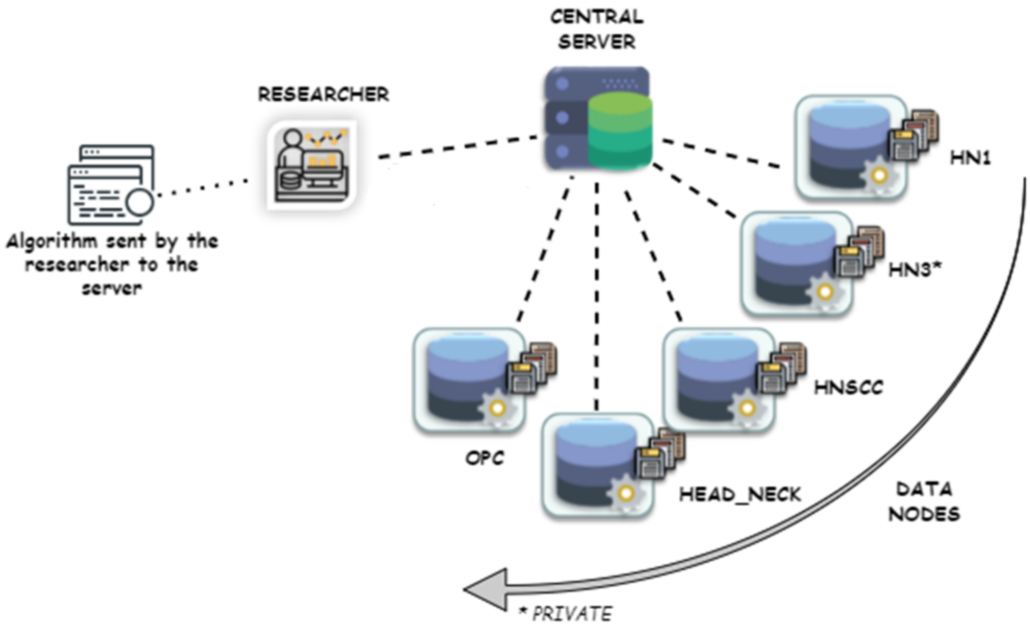
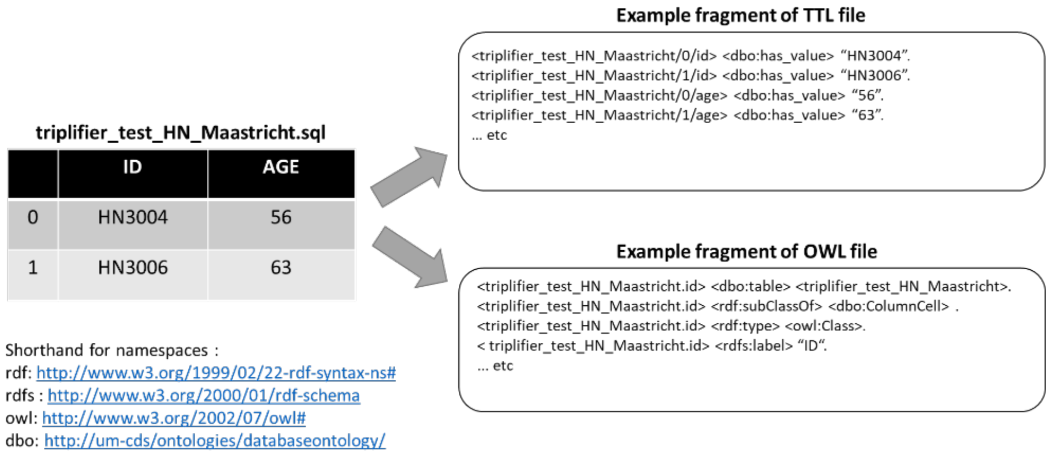


Figure 2: Schematic illustration of the Vantage6 infrastructure used in this work to show a likely clinical use case



triplifier_test_HN_Maastricht.sql

	ID	AGE
0	HN3004	56
1	HN3006	63

Shorthand for namespaces :
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs : http://www.w3.org/2000/01/rdf-schema
owl: http://www.w3.org/2002/07/owl#
dbo: http://um-cds/ontologies/databaseontology/

Example fragment of TTL file

```
<triplifier_test_HN_Maastricht/0/id> <dbo:has_value> "HN3004".
<triplifier_test_HN_Maastricht/1/id> <dbo:has_value> "HN3006".
<triplifier_test_HN_Maastricht/0/age> <dbo:has_value> "56".
<triplifier_test_HN_Maastricht/1/age> <dbo:has_value> "63".
... etc
```

Example fragment of OWL file

```
<triplifier_test_HN_Maastricht.id> <dbo:table> <triplifier_test_HN_Maastricht>.
<triplifier_test_HN_Maastricht.id> <rdf:subClassOf> <dbo:ColumnCell> .
<triplifier_test_HN_Maastricht.id> <rdf:type> <owl:Class>.
< triplifier_test_HN_Maastricht.id> <rdfs:label> "ID".
... etc
```

Figure 3: Example showing the expected output of triplifier processing. A hypothetical input table is shown on the left. Namespace aliases are used to improve readability. A fragment of the serialized database contents is shown top right in the TTL file, and a part of the database schema is shown with a database-specific ontology in the OWL file at bottom right

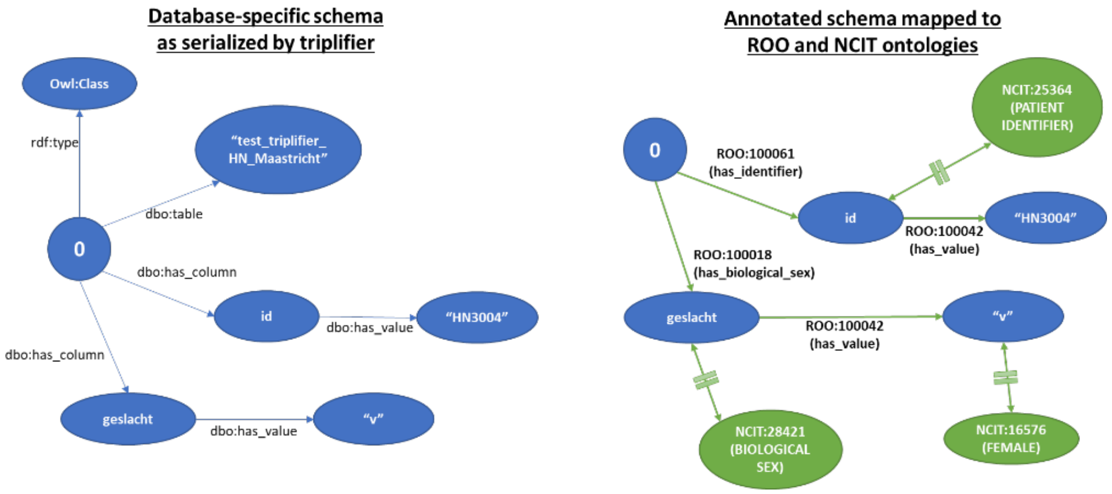
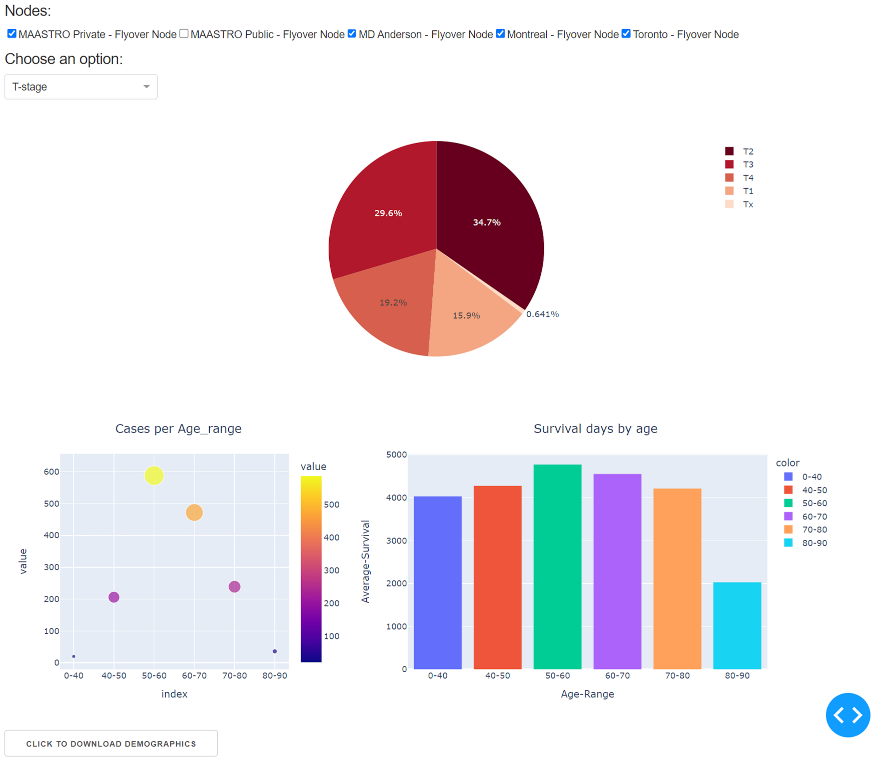


Figure 4: Examples showing knowledge graph of a patient "0" with classes ID and biological sex. The image on the left is from Triplifier after the data has been converted to RDF. On the right, is the image after the annotation graph has been added. Double-sided green arrows with double crossing bars are the predicate owl:equivalentClass



Nodes:

☑ MAASTRO Private - Flyover Node ☐ MAASTRO Public - Flyover Node ☑ MD Anderson - Flyover Node ☑ Montreal - Flyover Node ☑ Toronto - Flyover Node

Choose an option:

T-stage

CLICK TO DOWNLOAD DEMOGRAPHICS

Figure 5: Distributed Dashboard

References

[1] Kenneth Clark et al. “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository”. en. In: *Journal of Digital Imaging* 26.6 (Dec. 2013), pp. 1045–1057. issn: 1618-727X. doi: 10.1007/s10278-013-9622-7. url: https://doi.org/10.1007/s10278-013-9622-7.

[2] Office of the Commissioner. *Real-World Evidence*. en. Publisher: FDA. Oct. 2022. url: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

[3] *Data - W3C*. url: https://www.w3.org/standards/semanticweb/data.

[4] *FAIR-IFICATION OF STRUCTURED CLINICAL DATA | medRxiv*. url: https://www.medrxiv.org/content/10.1101/2021.07.23.21261032v3.full.

[5] Aaron Grossberg et al. *HNSCC*. Version Number: 2 Type: dataset. 2020. doi: 10.7937/K9/TCIA.2020.A8SH-7363. url: https://wiki.cancerimagingarchive.net/x/sIN5Ag.

[6] Jennifer Yin Yee Kwan et al. *Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma*. type: dataset. 2019. doi: 10.7937/TCIA.2019.8DHO2GLS. url: https://wiki.cancerimagingarchive.net/x/XAQGAg.

[7] Arturo Moncada-Torres et al. “VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange”. In: *AMIA Annual Symposium Proceedings* 2020 (Jan. 2021), pp. 870–877. issn: 1942-597X. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075508/.

[8] *Open Data and FAIR Data: differences and similarities | Plataforma OGoov*. en-US. May 2019. url: https://www.ogoov.com/en/blog/open-data-and-fair-data-differences-and-similarities/.

[9] *OWL - Semantic Web Standards*. url: https://www.w3.org/OWL/.

[10] *RDF - Semantic Web Standards*. url: https://www.w3.org/RDF/.

[11] Johan van Soest et al. “Annotation of existing databases using Semantic Web technologies: making data more FAIR”. en. In: (), p. 8.

[12] *SPARQL - Semantic Web Standards*. url: https://www.w3.org/2001/sw/wiki/SPARQL.

[13] *Syntactic and Semantic Interoperability | Electrosoft*. en. url: https://www.electrosoft-inc.com/resources/syntactic-and-semantic-interoperability.

[14] *The Personal Health Train Network | The Personal Health Train*. en. url: https://pht.health-ri.nl/personal-health-train-network.

[15] Stelios Theophanous et al. “Development and validation of prognostic models for anal cancer outcomes using distributed learning: protocol for the international multi-centre atomCAT2 study”. In: *Diagnostic and Prognostic Research* 6.1 (Aug. 2022), p. 14. issn: 2397-7523. doi: 10.1186/s41512-022-00128-8. url: https://doi.org/10.1186/s41512-022-00128-8.

[16] *Turtle - Terse RDF Triple Language*. url: https://www.w3.org/TeamSubmission/turtle/.

[17] Martin Vallières et al. *Data from Head-Neck-PET-CT*. type: dataset. 2017. doi: 10.7937/K9/TCIA.2017.8OJE5Q00. url: https://wiki.cancerimagingarchive.net/x/24pyAQ.

[18] Leonard Wee and Andre Dekker. *Data from Head-Neck-Radiomics-HN1*. type: dataset. 2019. doi: 10.7937/TCIA.2019.8KAP372N. url: https://wiki.cancerimagingarchive.net/x/iBglAw.

[19] *Welcome*. en. url: https://docs.vantage6.ai/.

[20] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". en. In: *Scientific Data* 3.1 (Mar. 2016). Number: 1 Publisher: Nature Publishing Group, p. 160018. issn: 2052-4463. doi: 10.1038/sdata.2016.18. url: https://www.nature.com/articles/sdata201618.

Supplementary

Table 1
Patient Demographics Table from five data nodes

	HN1	*HNSCC*	*OPC*	*HEAD-NECK*	*HN3*
Sample size	137	492	606	298	165
Age in years					
Mean	61.9	57.8	60.5	63.3	62.6
Range	44-83	28-87	33-89	18-90	29-84
Sex					
Female	26	69	125	71	43
Male	111	423	481	227	122
Tumour stage					
T1	35	92	103	39	14
T2	32	203	198	109	31
T3	24	117	183	94	68
T4	46	80	122	46	52
Tx	-	-	-	10	-
Nodal stage					
N0	60	45	101	59	48
N1	16	53	61	40	45
N2	58	378	397	180	54
N3	3	16	47	19	18
Nx	-	-	-	-	-
Metastasis stage					
M0	136	492	606	294	165
M1	1	0	-	0	0
Mx	-	-	-	4	-
Overall stage (7th ed.)					
I	24	3	11	4	-
II	11	16	38	27	-
III	23	67	85	61	-
IV‘	79	406	472	204	-
Unspecified	-	-	-	2	-
Tumour location					
Nasopharynx	-	-	-	28	-
Oropharynx	88	492	606	203	63
Hypopharynx	-	-	-	13	31
Larynx	49	-	-	45	64
Unknown	-	-	-	9	-
HPV status					
Positive	23	248	356	78	34
Negative	58	44	143	46	29
Unknown	56	200	107	174	102
Radiotherapy type					
Radiotherapy	100	57	309	48	104
Chemoradiotherapy	37	435	297	250	61
Survival status					
Censored	63	376	347	242	77
Deceased	74	116	259	56	88