City University of London

MSc Data Science

2019-2021

Project Report

A knowledge graph to orchestrate a multi-organ quantitative assessment of long Covid

Author: Judith Grieves

Supervised by: Ernesto Jiménez-Ruiz

Valentina Carapella (Perspectum)

Submitted: 5 December 2021

Declaration of Authorship

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed

Judith Grieves

Abstract

The quantitative measures produced from Magnetic Resonance Imaging clinical studies are an invaluable resource for analysis and future work but only if they are well-structured and easily accessible. This project investigates whether Semantic Web Technologies are a useful means of achieving these objectives. The project analyses data gathered for a single study on long Covid patients, models its meaning in an ontology and stores it in a knowledge graph. *SPARQL* queries are provided to extract subsets of the data for given parameters. We evaluate several current methods, tools and standards used to make the creation task easier, more efficient and repeatable: the Pay-as-you-go methodology, RDF Mapping Language and the *FAIR* data standards. The resulting proof of concept design and build successfully answers given business *competency questions* and should provide a useful starting point to take forward into building a comprehensive, flexible and useful tool.

Keywords: ontology, knowledge graph, magnetic resonance imaging, RDF, SPARQL

Acknowledgements

This work was conducted using the Protégé ontology editor, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

Table of Contents

Figures

Tables

1 Introduction and Objectives

Data generated by clinical studies is a valuable commodity for the scientists producing, storing and using it. Measurements of organ health from Magnetic Resonance Imaging (*MRI*) present a particular challenge as the quantitative metrics involved are important entities in their own right. Each metric may have many versions and requires its own attributes such as status and provenance. Organisations carrying out imaging clinical studies and wishing to access the data need an accurate and structured way to consolidate and query the results. Semantic Web Technologies (SWT) provide a flexible and knowledge-based means of holding and accessing data and may be a good solution to this problem.

In conjunction with Perspectum[1], this project will design and build software to store data from imaging clinical studies, using their recently published paper of long Covid patients (Dennis et al, 2020), investigating the extent of organ damage in affected participants. This will serve as a proof of concept for their wider study results and the evaluation of the project will determine whether these technologies provide a useful solution for Perspectum's requirements.

1.1 Background

1.1.1 Perspectum

Perspectum is an Oxford-based medical imaging company, founded in 2012 and specialising in non-invasive multi-organ imaging to assess patient health. Health assessment metrics from *MRI* have been shown to serve as reliable proxies for a number of *biomarkers*, measurements of organ health, previously only obtained by invasive biopsies (Hutton et al., 2018). Perspectum products include LiverMultiScan[2], to support liver diagnostics, and Atlas[3], a multi-organ assessment of images of liver, spleen, pancreas, kidneys, heart and aorta.

As part of their research and development, Perspectum scientists carry out clinical studies that collect data consisting of imaging metrics and other measurements relating to patients and their health.

Direct (or raw) study data is stored in an internal system to a standard required both for clinical and Contract Research Organisation (CRO) customers. The derived data about these metrics is currently held in multiple spreadsheets and is not easily accessible. Data Scientists are currently responsible for the creation of validated metrics, manipulating and analysing this data to respond to information requests from others within the organisation. These data analysis tasks can be time-consuming and there is a requirement that the data is held centrally, accessible and structured in a way that makes the process more efficient and could even allow non-experts to formulate and answer queries.

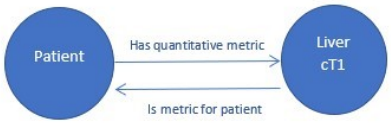
[1] https://perspectum.com/ (accessed 28/9/2021)
[2] https://perspectum.com/products/livermultiscan (accessed 28/9/2021)
[3] https://perspectum.com/products/atlas (accessed 28/9/2021)

1.1.2 Semantic Web Technologies

The core of Semantic Web Technologies are ontologies and knowledge graphs. An ontology traditionally models the abstract knowledge and structure of a domain as concepts, relationships and axioms, whilst the knowledge graph contains instances of the concepts from the domain. The entities and relationships of Ontologies and knowledge graphs can be represented as directed, labelled graphs, a mathematical concept consisting of vertices (or nodes) and edges (lines) between them (Figure 1a).

Graphs can provide a useful abstraction to represent real world concepts and their relationships to one another. They can be represented by 'triples' of subject, predicate and object showing that one node 'has some relationship to' another node (Figure 1b).

The Resource Definition Framework (RDF)[4] is a standard way of representing and storing these triples (Figure 1c), where each element of a triple is considered a Resource, with a Uniform Resource Identifier (*URI*). *RDF* is supported by the World Wide Web Consortium (W3C)[5], the international Semantic Web standards community. As well as representing the knowledge in an ontology, *RDF* triples can also hold the instances of data in a knowledge graph (Figure 1d).



Subject	Predicate	Object
Patient	has patient metric	Liver cT1
Liver cT1	is metric for patient	Patient

```
@prefix : <http://example.org/resource/> .

:Patient      :hasPatientMetric    :LivercT1
:LivercT1     :isMetricForPatient :Patient
```

Subject	Predicate	Object
URI_liver_cT1_P001	isMetricForPatient	URI_Patient_P001
URI_liver_cT1_P002	isMetricForPatient	URI_Patient_P002
URI_Patient_P001	hasPatientMetric	URI_liver_cT1_P001
URI_Patient_P002	hasPatientMetric	URI_liver_cT1_P002

Figure 1. (a) top left: concepts represented in a graph. (b) top right: relationships as triples. (c) bottom,left: triples as RDF. (d) bottom right: instances of triples.

RDF forms the basis of the W3C technology stack which also includes OWL[6], a language that allows logic and inferencing around these concepts and *SPARQL*[7], a declarative RDF query language specifying the required output, ordering and route over the graph data.

Semantic Web Technologies can be a key component for Artificial Intelligence (AI), Machine Learning (ML) and the digital processing of large amounts of data. In AI and ML, machines do not

[4] https://www.w3.org/RDF/ (accessed 29/9/2021)
[5] https://www.w3.org/ (accessed 4/10/2021)
[6] https://www.w3.org/OWL/ (accessed 4/10/2021)
[7] https://www.w3.org/TR/rdf-sparql-query/ (accessed 5/10/2021)

'understand' the data (Shimizu, 2019) but if it is properly structured, Ontologies can 'bridge the gap' to bring meaning and understanding.

1.2 Aims and Objectives

The aim of this project is to create a set of digital products, as a proof of concept, to meet the business requirements of Perspectum whilst also assessing and reporting on the value of recent academic methods and tools applicable to the solution.

1.2.1 Objectives

More specifically, our objectives are:

A. To create an ontology, a systematic, formal, and unambiguous representation of the knowledge around imaging and other patient health measurements for the study data, characterising the profile of these values, that is, the data needed to unambiguously identify each one.
B. To (automatically) create a knowledge graph containing instances of these concepts loaded from tabular input data.
C. To research and evaluate current methods of creating this knowledge data that are robust, repeatable, and accurate.
D. To investigate current methods of accessing semantic data so that users can benefit from the querying capabilities of knowledge graphs.

1.2.2 Research Questions

The research questions this study will aim to answer are:

1. Can ontologies provide a useful unified view of the complex use case presented by the Perspectum study data?
2. Can the data in this domain, that is, measurements related to liver imaging clinical studies, be put into context using a knowledge graph? Are the hierarchies and relationships at the knowledge graph level useful to profile this data?
3. Can we (semi)automate and streamline the creation of the ontology and knowledge graph via templates and other recently researched methods?
4. What features of ontologies and knowledge graphs recommended in current literature should be included to enhance the usefulness of the products?
5. Is there a feasible alternative to the creation of a new Semantic data source, for example, the use of an ontology and mappings to access the data in a non-Semantic form or the use of alternative database technologies such as relational?

1.2.3 Beneficiaries

Perspectum Research & Development: this study will evaluate the benefits or otherwise of representing and analysing this domain of data with Semantic Web Technologies. It aims to create a

prototype ontology and knowledge graph and demonstrate useful means of querying and accessing the data. Perspectum will have the design and build products and this report of the options available to explore these methods further if the benefits of doing so are sufficient.

City University[8] Research: this study will add to the literature on the domains that can benefit from Semantic Web Technologies and provide documentation of the strengths and weaknesses found in the methods and tools used in the project.

The researcher: this study will allow me to gain experience of an academic research project and in formulating a Semantic Web solution to a real-world problem, exploring the options available.

1.2.4 Evaluation Criteria

1. *Competency questions* (Bezerra et al, 2013). Perspectum will supply business questions that should be answered by the knowledge graph. We will describe how well these can be answered by the software created.
2. Evaluations of tools and methods will specify individual evaluation criteria.
3. *Findable, Accessible, Interoperable and Reusable (FAIR)* principles (Wilkinson, 2016) will be used to assess how well the ontology and knowledge graph meet current standards of semantic data.

1.3 Methods and Work Plan

This work will be carried out as a design and build project.

The clinical data collected in the long Covid study (Dennis et al, 2020) will be used to evaluate the usefulness of ontologies and knowledge graphs for this particular biomedical domain.

1.3.1 Methods

As input, Perspectum will provide a collection of synthesised, tabular datasets, representing a subset of the study data and based on reasonable, realistic patient scenarios from the published research.

These datasets, as well as the input of domain experts, will be used to model the domain as an ontology, providing a structure for a resulting knowledge graph.

Development will be organised in short iterations. A limited scope of data items will be used in early iterations, to create the proof of concept and allow a base from which to expand when appropriate. A subset of data relating to a single organ, the liver, will be modelled.

[8] https://www.city.ac.uk/about/schools/mathematics-computer-science-engineering/computer-science (accessed 28/9/2021)

1.3.2 Work plan

A literature review is carried out to research and assess the current state of Semantic Web Technologies, the latest proposed tools and methods and any known ontologies or knowledge graphs related to this domain.

Using knowledge gained in this review and following the Pay-as-you-go ontology project methodology (*PAYG*) (Sequeda, 2019), which we will use and evaluate, multiple short iterations were planned (Appendix A), each comprising:

- analysis of the data and gathering of requirements in conjunction with domain experts (where appropriate).
- modelling, design and creation of an ontology: using the results of the data analysis.
- design, development and testing of the data migration software to create the knowledge graph: using the ontology structure and input data format.
- creating queries to question the knowledge graph and provide required results.

These tasks follow the *PAYG* high-level stages of Knowledge Capture, Knowledge Representation and Self-Service Analytics. Figure 2 shows these stages and the deliverables to be created in each.

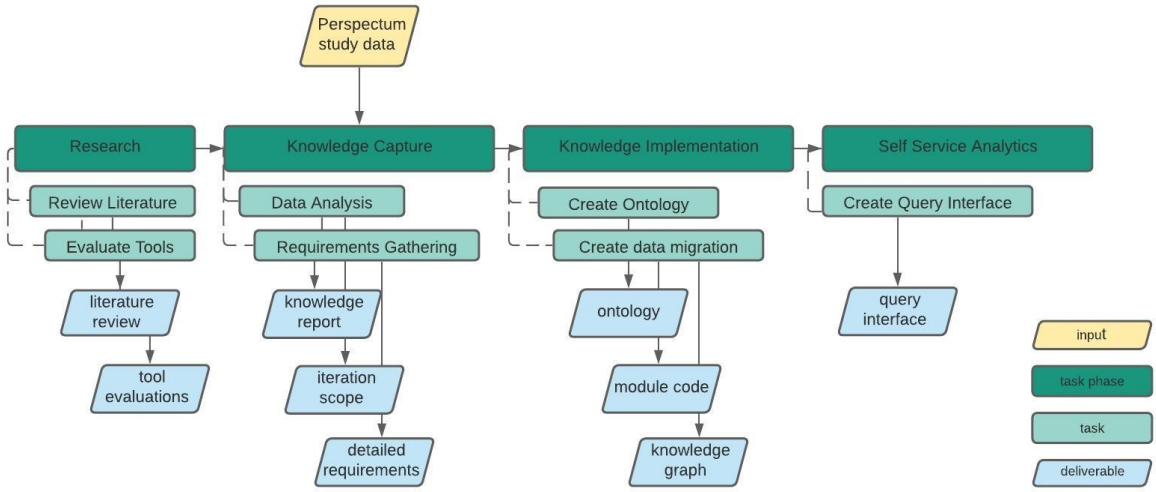


Figure 2. Project tasks and deliverables

1.3.3 Products Created

The major outputs of this project are:

1. Documented requirements of the project[9].
2. Ontology of the domain knowledge in scope[10].

[9] KnowledgeReport.xlsx - Google Sheets (accessed 3/12/2021)
[10] City-MSc-Project/ontology at main · JudithGrieves/City-MSc-Project (github.com) (accessed 3/12/2021)

3. A set of software modules to take the supplied tabular data and create as a knowledge graph[11].
4. A knowledge graph of the supplied data[12].
5. A means of accessing and interrogating the created knowledge graph to allow data querying, extraction and analysis[11].
6. Evaluations of the tools and methods used.
7. Presentation to Perspectum staff (slides 11-18)[13]

1.3.4 Tools Used

The following proprietary and open software tools are used in the project:

- Requirements' documentation: MS Word, MS Excel, Google docs, Google sheets
- Ontology modelling: the Protégé[14] ontology modelling tool (Musen, 2015)
- Transformation and load: Python (and any other evaluated tools, detailed further below)
- Querying and testing: *SPARQL* (query language) executed from custom Python modules.

1.3.5 Changes to the Proposal Project plans

The original project proposal work plan (Appendix A) showed a 2-week literature review and a 2-week tools evaluation, followed by 3 month-long iterations of a build cycle, each comprising analysis, design, build, test and evaluation. Once the project began, it became clear that it was more flexible to carry out the tools evaluations within iterations and interleave them with the analysis-design-build cycles. The (co-)supervisors were available to meet fortnightly so the work was divided into 2-week long iterations. In this way, each iteration would end with a meeting where we reviewed the completed work and agreed the contents of the next.

1.4 Structure of the Report

Section 2, Context, documents the results of a literature review into the history and current state of Semantic Web research, the use of ontologies and knowledge graphs in general and in related biomedical domains, and the latest tools and methods used in their creation.

Section 3 describes the methods used in the project, that is, the project organisation and communication, the iterative working, the stages of the development methodology used and the testing and evaluation of outputs.

---

[11] City-MSc-Project/code at main · JudithGrieves/City-MSc-Project (github.com) (accessed 3/12/2021)
[12] City-MSc-Project/data at main · JudithGrieves/City-MSc-Project (github.com) (accessed 3/12/2021)
[13] https://docs.google.com/presentation/d/1GoRC7GKgadbaMNHGIlAnj2HHTnnmCJQS1CEM402QUo0/edit#slide=id.p (accessed 3/12/2021)
[14] https://protege.stanford.edu/ (accessed 28/9/2021)

Section 0 documents the results of the design and build, the iteration contents, descriptions and examples of the software created, and the results of the evaluations of tools and methods carried out.

Section 5 discusses the implications of the results obtained and how far they answered the research questions and met the project objectives.

Section 6 evaluates the overall project, identifies particularly positive or negative results and highlights areas of further study.

Glossary contains a glossary of terms used in this report. Glossary terms in the report are given in *italics*.

Appendices contain detailed versions of work referred to in the report:

Appendix A : RMPI Project Proposal for MSc in Data Science

Appendix B : Revised Project Plan

Appendix C : Knowledge Report

Appendix D : Evaluation – FAIR data

Appendix E : Python Code

Appendix F : RML Mapping File

Appendix G : Ontology

Appendix H : Test input data

Appendix I : Output Query Test Results

2 Context

2.1 Semantic Web Technologies

In 2001 Tim Berners-Lee defined a vision of the Semantic Web (Berners-Lee 2001) as an extension of the current web, an interlinked network of structured data that could be accessed and read by machines performing tasks for users. The data would encompass knowledge representation in Ontologies —collections of information defining the relations between concepts— with rich metadata and inference rules. He positioned the *Resource Definition Language (RDF)* as a standard means of abstracting this graphical data of nodes (entities) and edges (relationships), as *triples* of subject, verb and object, where all concepts would be uniquely identified by a *Universal Resource Identifier (URI)*. Ontologies are defined in the *W3C* Web Ontology Language (*OWL*) and data is extracted using the *SPARQL* query language.

He also defined a straightforward set of guidelines (Berners-Lee 2006) to be met, for linked data to be considered '5 star' in terms of availability, readability, non-proprietary formatting, open and linked. The guidelines recommend that data should be available on the Web, identified by *HTTP URIs*, readable by machines using the *RDF* standards and linked to *URIs* of other data.

Since this beginning, the research field has expanded to encompass a wider interest in using the Semantic Web methods and tools (Hitzler 2021a) for other, non-Web, uses of 'data sharing, discovery, integration and re-use' and to demonstrate that ontologies, linked data and knowledge graphs are useful for more general means of data management.

Although the original vision of the Semantic Web has not yet been widely realised (Hogan, 2019), there are examples of the successful use of Semantic technologies on the Web, e.g., Wikidata, DBpedia, and elsewhere in a wide range of other domains, as diverse as biomedicine (Smedley, D. et al, 2021), offshore oil platform building (Skjæveland et al, 2018b) and digital humanities (Shimizu et al, 2020).

2.2 Biomedical domain

Knowledge representation has been widely used in the biomedical domain for over 20 years. An ontology's vocabulary, metadata and machine-readable axioms (Hoehndorf, R et al.,2015) are particularly useful to characterise the hierarchies and taxonomies found in biomedicine.

As a result, there are many well-known and well-used biomedical ontologies. One of the earliest and best known, the Gene Ontology[15] (Ashburner, M. *et al.*, 2000) (The Gene Ontology Consortium *et al.*, 2021), was created in 1998 (Carbon 2018) and is widely used in the life sciences. It is continually

[15] http://geneontology.org/ (accessed 28/9/2021)

updated and currently contains the results of over 150,000 published papers in 700,000 evidenced annotations.

Discovery and reuse of existing ontologies is encouraged and facilitated by the many portals cataloguing biomedical ontologies. The National Center for Biomedical Ontology (NCBO) (Musen et al, 2012) supports biomedical research by providing BioPortal[16] (Whetzel et al, 2011), a collection of ontologies that follow NCBO-recommended formats and methodologies. The Open Biological and Biomedical Ontologies (OBO) Foundry[17] exists to co-ordinate the accurate evolution of ontologies to support biomedical data integration (The OBO Foundry, 2021); their ontologies and principles provide useful guidance.

Biomedical knowledge graphs have been put to many uses, retrieving large amounts of data, providing supporting knowledge, inferring new hypotheses and simplifying representation of complicated data. A recent review of 2018-19 papers (Callahan et al, 2020) found 83 covering knowledge graphs and biomedical subjects, ranging from genomic data to clinical decision support.

2.2.1 Ontologies of Liver and Imaging Data

The project domain encompasses medical imaging, liver anatomy and diseases and clinical study results. A recent systematic literature review of ontologies of imaging and liver disease (Messaoudi 2020) found 39 relevant studies between 2005-19. Most of these studies applied ontologies as decision support in the diagnosis of disease (Wahab, 2019), often in the interpretation of images or disease representation.

Ontologies have been used in processing liver imaging data to detect and stage tumours (Messaoudi et al, 2019a), showing reliable results in real-world case studies (Figure 3). Similarly, hepatocellular carcinoma (HCC) was successfully diagnosed and staged (Messaoudi, 2019b) using an ontology created from extracts of *magnetic resonance imaging (MRI)* report data, liver lesion medical reports and reasoning in Protégé[14].

As early as 2007, semantic technologies were being used to model the radiographic domain, with the use of Protégé[14] to create ontologies of medical imaging findings and interpretations as subsets: patient observations, the imaging domain and an anatomy reference (Marwede et al, 2007).

Ontologies are also used to model Radiological Reports, primarily in the diagnosis of disease by identifying relevant text in the reports. Unstructured natural language text can be parsed to extract liver image diagnostic data (Kökciyan, 2014). In this case, the semantic inferencing allowed by

[16] https://bioportal.bioontology.org/ontologies (accessed 28/9/2021)
[17] http://www.obofoundry.org (accessed 23/4/2021)

ontologies, in combination with the standard radiology vocabulary *RadLex*[18], can extract meaning from the reports without the inconsistencies and ambiguities of the original text.

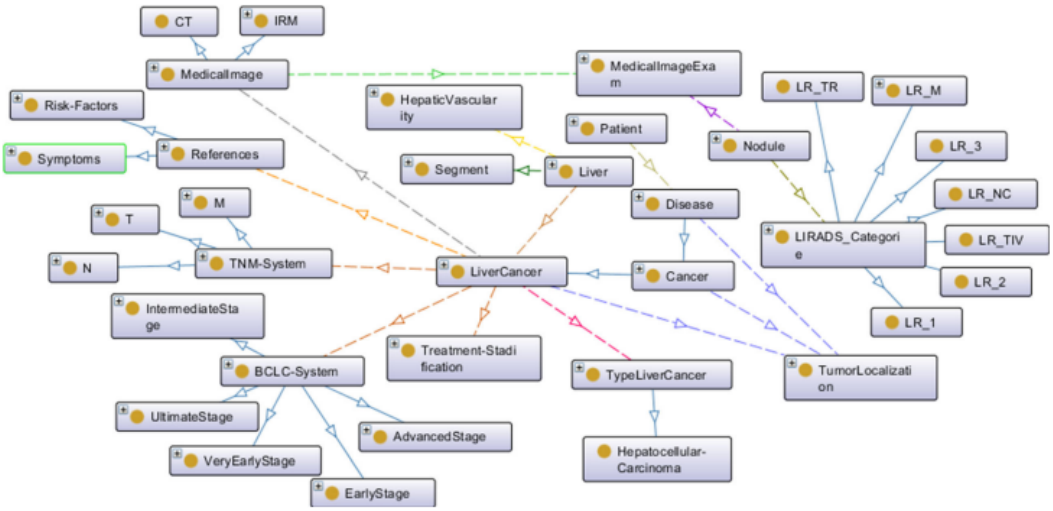


Figure 3. Part of the liver disease ontology (Messaoudi, R. et al. 2019a)

Radiological reports have also been used to assess the quality of images from the UK Biobank[19] (Carapella et al, 2016), where free text report data is extracted and modelled as an ontology of heart image quality. This ontology, with entities of Imaging_Scan_Visit, Medical_Imaging_Study, personId, participantId and scandate, could be usefully reused or referenced in this project's ontology. Linking the project ontology to these overlapping entities would illustrate the important benefits of Semantic Web, that is, leveraging existing work in a related field. A 2018 study covering liver disease patient cases (Roldan-Garcia, 2018) created an ontology, LiCO[20] (Liver Case Ontology), believed at the time to be the first case-centric ontology. The model is based on patient cases, using liver disease data to create a proof of concept and consists of classes of, for example, Patient, Study

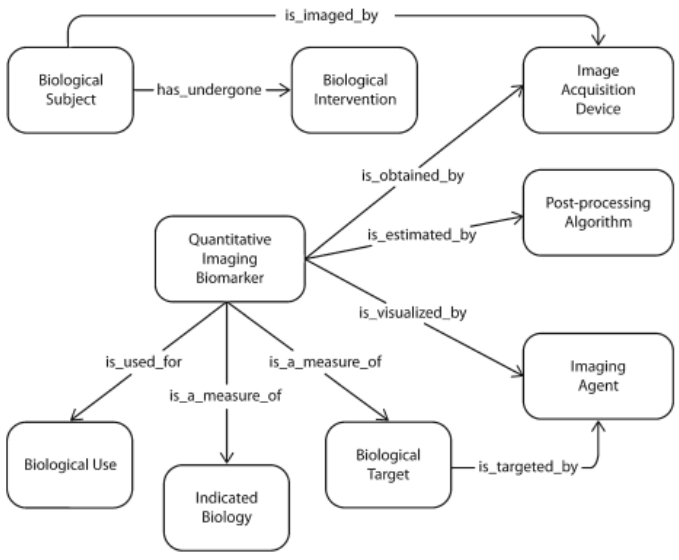


Figure 4. QIBO top level classes and relationships (Buckler et al., 2013)

[18] http://radlex.org/ (accessed 28/9/2021)
[19] https://www.ukbiobank.ac.uk/ (accessed 30/9/2021)
[20] https://bioportal.bioontology.org/ontologies/LICO (accessed 30/9/2021)

and other properties used in the Perspectum case. The paper highlights how aligning to an existing ontology, ONLIRA[21] (Kokciyan et al., 2014), and existing vocabularies, in this case *SNOMED CT*[22] and *RadLex*[18], can be used to improve query results.

The earliest ontology found to model imaging *biomarkers* created the Quantitative Imaging Biomarker Ontology (QIBO) (Buckler et al., 2013), motivated by the wish to enhance the use of available imaging data. Relevant terms were collected from studies in the journal Molecular Imaging and Biology and then modelled (Figure 4) based on the structures of various publicly available ontologies. The resulting ontology was able to find potential new *biomarkers* for a pair of diseases.

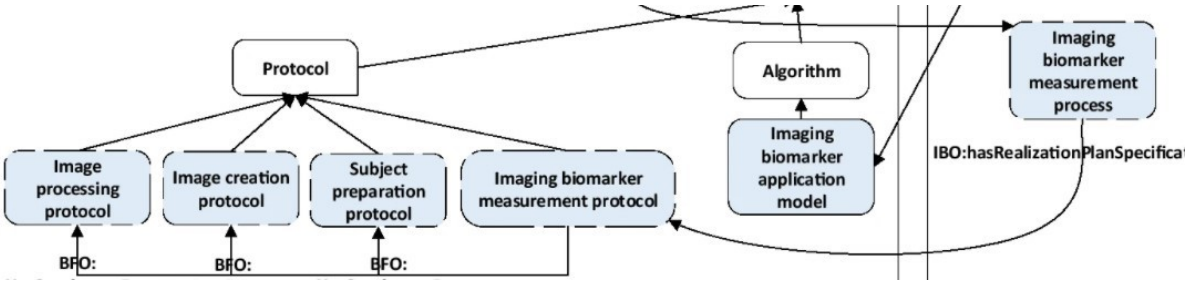


Figure 5. IBO modelling of the provenance of an imaging biomarker (Amdouni et al., 2018).

A further work, creating the IBO ontology, to model quantitative *biomarkers* also has significant overlap with this project's data (Amdouni et al., 2018). This highlights the importance of these measurements as *first-class entities* and the critical nature of correct modelling to link them to their study, measurement protocol (Figure 5) and other provenance data to facilitate sharing and reuse of study data. Modelling this ontology emphasised the use of a foundational ontology (BFO), reuse of well-known existing ontologies, eg. OBI, FMA, GO, and the use of modular methods to aid modelling and reuse.

IBO and QIBO look like very useful ontologies to provide input to this project's modelling but unfortunately neither are currently available online. The literature search did not find any other examples close enough to this project's specific use case and granularity to reuse wholly but we will look for opportunities to link individual entities in our newly created ontology to existing ones where overlap is found, for example re-using *URI*s or creating synonyms to external resources.

2.2.2 Clinical Study ontologies

There has been ongoing use of ontologies and knowledge modelling in the clinical research domain. A special issue of the Journal of Biomedical Informatics on ontologies for clinical and translational research described common themes for the future of ontology development and use, as well as the

[21] https://bioportal.bioontology.org/ontologies/ONLIRA (accessed 30/9/2021)
[22] https://www.snomed.org/ (accessed 28/9/2021)

ontology frameworks and principles that emerge from the papers covered in the issue (Smith et al, 2011).

Ontologies to model and store the results of clinical studies include The Ontology of Clinical Research (OCRe) (Tu, 2009) and, more recently, the Ontology-Based eXtensible data model (OBX) (Kong et al, 2011). OBX is a comprehensive ontology for The U.S. National Institutes of Health, based on reference ontologies including the Basic Formal Ontology (BFO)[23] and the Ontology for Biomedical Investigation (OBI)[24]. OCRe and OBX have some overlap with this project's scope in the classes of Study, Event (Visit) and Value (Quantitative Metric) but are much larger supersets, encompassing data required for the full lifecycle of a complex clinical study rather than the end results in which we are interested.

2.3 Standards, Tools and Methodologies

The Semantic Web field is widely believed to be suffering from an abundance of inconsistent data, methods and tools and requires consolidation that is likely to be driven by industrial requirements backed up by academia (Hitzler, 2021a). We discuss some of the recent methodologies, tools and standards suggested in the literature.

2.3.1 Methodologies

Ontologies are still considered difficult to deploy and require highly qualified practitioners so there is a need for development methodologies to simplify the process. A recent Pay-as-you-go methodology (PAYG) takes an iterative approach to building ontologies and knowledge graphs from existing databases (Sequeda, 2019), breaking the project into iterations of data subsets to answer specified business *competency questions*. This methodology requires only a minimum of knowledge of OWL logic as it finds a small subset of OWL expressivity is sufficient for business problems encountered. To clarify discussions with domain experts not familiar with ontology modelling, it uses a vocabulary of Concepts, Attributes and Relationships as preferable synonyms to the OWL/RDF terms of Class, Data Property and Object Property.

Although this methodology is defined for relational database sources and this project's source datasets are *CSV* files, the structure and tasks are transferable and should provide a useful framework and workflow.

2.3.2 Ontology Development

An alternative to modelling large existing databases on an iterative 'pay-as-you-go' basis is to automate the process (Jiménez-Ruiz et al, 2015), employing a tool to extract a preliminary ontology from a database schema. The tools follow a general rule of translating tables to classes, *Foreign Key*

[23] https://basic-formal-ontology.org/ (accessed 28/9/2021)
[24] http://obi-ontology.org/ (accessed 28/9/2021)

attributes to *object properties* and other attributes to *data properties*. The provisional ontology can then be validated and amended before being used for mappings to the original data.

A current issue in the SWT community is concerned with the proliferation and overlap of ontologies (Smith et al, 2011). Re-use of ontologies is an important facet of Semantic Technologies and modular development has been explored as a means of encouraging this. Existing ontology tools are low-level without any means of abstraction and can therefore be 'inefficient, repetitive and error prone' (Skjæveland et al, 2018a).

MODL is a modular ontology design library that uses modular development and reuse of design patterns from multiple domains (Shimizu et al, 2019). Particularly emphasised is the use of graphical schema diagrams to elicit knowledge from domain experts and the use of concepts and relationships meaningful to them, as well as following established modelling principles (Figure 6). Templates are suggested to simplify ontology development.

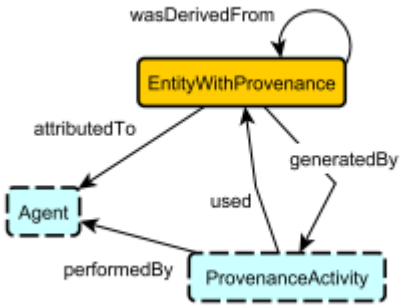


Figure 6. Example of a MODL Ontology Design Pattern (Cogan et al, 2020)

Modular Ontology Modelling (MOMo) is a modular development method (Shimizu et al, 2021), with a focus on the division of the ontology into modules and an emphasis on reusing design patterns from existing libraries such as MODL. MOMo aims to address good modelling principles and the 'failure' of the desired ontology reuse because of mismatched granularity and lack of clarity to domain experts. The MOMo workflow consists of 10 steps from describing use cases, data sources and *competency questions* through to creation of OWL files, giving detailed guidance on each step and with an emphasis on the importance of documentation in an ontology.

Extracting repeated patterns in the knowledge structure to simplify the creation and maintenance is also at the core of the OTTR[25] toolset, a language and template-based means of creating ontologies (Skjæveland et al, 2018a). OTTR aims to hide the complexity of ontology structures, create more uniform models and avoid omissions where structures are repeated.

Whilst modularisation and templates streamline the creation of an ontology there is a lack of independent tools allowing multiple ontology users to collaborate in editing, publishing and reviewing. The integration of WebProtege[26] and BioPortal[16] allows collaboration of multiple

[25] https://www.ottr.xyz/ (accessed 30/9/2021)
[26] https://webprotege.stanford.edu/ (accessed 28/9/2021)

designers, especially useful in large unwieldy ontologies, by using simultaneous editing to allow sharing of feedback and discussion from designers and users (Noy et al, 2010).

Similarly, ROBOT (Jackson 2019) is an open-source library and process tool to automate the related ontology development tasks of requirements gathering, ontology development and publishing, feedback, and deployment. It also performs quality control and checks ontologies for certain logical errors.

2.3.3 Knowledge Graph Creation

The development of large, complex knowledge graphs can be difficult and requires tools and methodologies to increase efficiency, accuracy and timeliness.

Creating knowledge graphs frequently involves migrating data from an existing source and format, so tools to support this stage are particularly important. There was no multi-source, multi-format tool available to carry out this migration using a mapping language until the 2013 proposal of RDF mapping language (*RML*) (Dimou, 2013).

RML[27] is an extension of the *W3C* reference language *R2RML*[28]—a mapping language from relational database to RDF—and supports multiple input formats, including *CSV*, *XML* or *JSON*, within a single mapping file (Dimou, 2014). *RML* consists of *RDF* statements to create a mapping between the source dataset and the desired *RDF* triples. As the mapping itself can be written in human-readable *Turtle* format RDF, it can also serve as a useful information document about the conversion process.

R2RML and *RML* have become the most prevalent method of knowledge graph creation. Note that the generation of the knowledge graph is carried out without any direct reference to an underlying schema or ontology. This means that there is no guarantee that the knowledge graph will correctly match the ontology though solutions exist to detect any inconsistencies in the mappings (Dimou, 2020).

There are several tools to execute *RML* mappings and create the knowledge graphs, including RMLMapper[29], which compared well for features and conformance to specification in an evaluation of available processors (Arenas-Guerrero et al, 2021).

As an alternative to the creation of new knowledge graphs, recent papers (Sequeda et al, 2019) (Kharlamov et al, 2015) describe Ontology-Based Database Access (OBDA) as a means of accessing existing data in non-Semantic structures by means of an ontology and mappings to the underlying

[27] https://rml.io/ (accessed 29/9/2021)
[28] https://www.w3.org/TR/r2rml/ (accessed 30/9/2021)
[29] https://github.com/RMLio/RML-Mapper (accessed 28/9/2021)

sources. The underlying database is exposed and presented as a virtual knowledge graph which users can query, using tools such as Ontop (Xiao G. et al, 2020).

2.3.4 Standards

Guidance to address the availability and reuse of data were published (Wilkinson 2016) as the *FAIR* principles—Findability, Accessibility, Interoperability, and Reusability. The principles aim to promote the goal of digital research objects being more easily 'found, re-used and cited over time' by human and machine users.

The *FAIR* principles are supported and propagated by a set of initiatives, GO-FAIR[30], that advocate by fostering cultural change, providing training and creating standards and infrastructure to support their implementation in digital research data. Another researcher resource to aid discovery and use of databases, standards and policies to help data products conform to the *FAIR* principles is FAIRSharing[31] (Sansone, S.-A et al., 2019) which, for example, lists the QUDT unit of measure vocabulary used in the project (FAIRsharing: QUDT, 2021b).

To mitigate against inconsistent implementations of these high-level principles, more prescriptive guidelines are given (Jacobsen et al, 2020b) and to aid the process of making data *FAIR*, a step by step, domain-agnostic workflow is defined (Jacobsen et al, 2020a).

Although the data in this project is intended to be used internally by Perspectum and not made publicly available, these guidelines and processes will prove useful to ensure any triples created conform to a set of standards and link to existing ontologies and knowledge graphs, to ensure domain validation and any future publication requirements, as well as applying the FAIR principles within the company.

2.4 Conclusion

Semantic web technologies are finding an ever-wider use in biomedical science and business and there are a number of initiatives in the research to make their development more accurate and robust. Even within organisations, creating ontologies for private use, an important principle is to follow widely adopted standards, reuse existing knowledge structures or provide connections to them. This work will adhere to these principles as far as possible.

[30] https://www.go-fair.org/fair-principles/ (accessed 17/8/2021)
[31] https://www.fairsharing.org/ (accessed 17/8/2021)

3 Methods

This chapter describes the activities carried out during this design and build project and how they were structured. For project planning and management, it includes the elements of the Pay-as-you-go ontology development methodology used (Sequeda, 2019), the breakdown of the tasks into short iterations and more general project management approaches. Also described are the methods of testing and validation of the products produced, the implementation of the software and the evaluation of the tools and methods used from the literature.

3.1 PAYG Methodology

Following the Pay-as-you-go methodology (Sequeda, 2019), software development tasks were divided into three main phases: Knowledge Capture, Knowledge Implementation and Self-service Analytics.

3.1.1 Knowledge Capture

Project requirements and data analysis tasks fall into the Knowledge Capture Phase: Analyse processes, collect documentation and develop a knowledge report.

Analysing processes included documenting the answers to the prescribed high level requirements analysis questions of What, Why, How, Where, When, Who and the population of a Knowledge Report (Appendix A.a.i.1.a.Appendix C) containing descriptions and details of the Concepts, Relationships and Attributes for the iteration.

Sources of information for the documentation collection phase came from:

- Domain expert meetings and interviews.
- Domain expert supplied information – data, background notes, etc.[32]
- Perspectum publicly available product information [33] [34] [35].

The knowledge report was created as pages in a spreadsheet.

3.1.2 Knowledge Implementation

The knowledge implementation phase encompasses the creation (or extension) of the ontology (objective A) with new data structures, the mapping and transformation of source data to the ontology entities (objectives B and C) and the extraction of answers to the business questions via *SPARQL* queries (objective D). The results of the queries are evaluated against expectations to determine their

[32] https://drive.google.com/drive/folders/1ZVCZWxNUCG7LpMhRV9-NOTLpxN0p-NKX (accessed 2/11/2021)
[33] https://perspectum.com/media/2081/mkt0073-ct1-explainer-40.pdf (accessed 4/10/2021)
[34] https://perspectum.com/media/2084/mkt0091-t2star-explainer-20.pdf (accessed 4/10/2021)
[35] https://perspectum.com/media/2083/mkt0089-pdff-explainer-20.pdf (accessed 4/10/2021)

success. The *PAYG* methodology specifies the following sub-areas of the Knowledge Implementation phase which were followed:

A. Create/Extend Ontology

Collating the sources of information, and understanding how data items relate to one another, allows the creation of the ontology entities: classes (concepts), *object properties* (relationships) and *data properties* (Attributes). For example:

- MRIScannerModel and ScannerFieldStrength are defined as ontology OWL classes
- 'hasScannerFieldStrength' is an object property connecting classes MRIScannerModel ScannerFieldStrength
- FieldStrengthValue is a data property of OWL class ScannerFieldStrength and has the *range* numeric 'double'

Ontology design was carried out in Protégé[14], a graphical tool to create OWL ontologies. Figure 7 shows examples of the graphical and export elements of the tool.

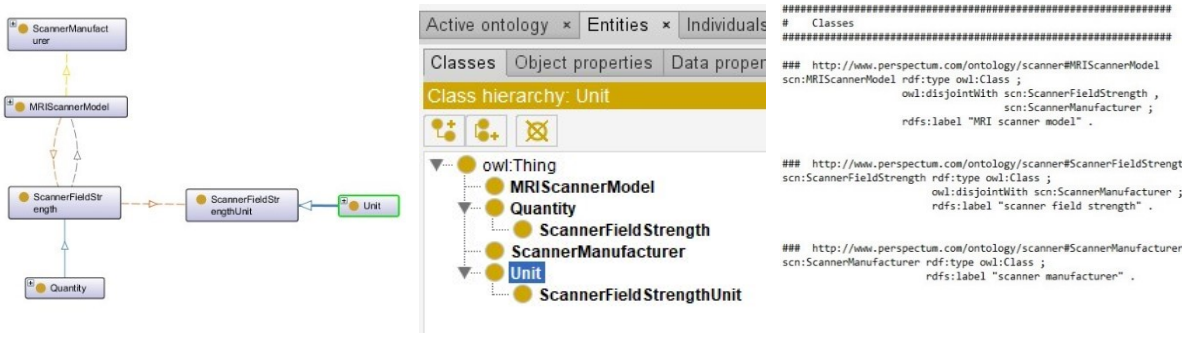


Figure 7. Protégé Tool elements Left: ontology diagram. Centre: Class definitions. Right: export Turtle format file.

To simplify the modelling, a subset of the OWL relationships' expressivity was used: classes, subClassOf, disjointWith, object property SubProperty, InverseOf, *Domains* and *Ranges*. Relationships are defined in both directions using the inverse property, eg.

- ScannerManufacturer isMakerOfScanner MRIScannerModel
- MRIScannerModel isMadeBy ScannerManufacturer
- isMakerOfScanner isInverseOf isMadeBy.

Domains and *ranges* define the classes pertinent to a relationship and were specified on only one side of the relationship, assuming the inverse would be inferred when querying the graph.

An important element of semantic modelling is to increase the interoperability of the ontology with links to others. The project ontology classes were linked to existing ontologies found in the surveyed literature or in a search of BioPortal.

B. *Data Transformation Design and Build (implement mapping)*

The data transform and load software is built to create a knowledge graph of *RDF* triples, from input *CSV* files.

Early iterations consisted of custom Python code written to execute this transformation, described in more detail in Iteration 1 (4.2.1) and

Iteration 2 (0).

Iteration 3 (section 4.2.3) evaluated the RDF Mapping Language (*RML*) and related tools, to carry out the data transformation and load. The use of RML required the creation of a new, but more straightforward, Python module to execute the open source RMLmapper[29], the tool selected to execute the transformation.

The evaluation of RML is discussed in more detail in section 4.7.

C. *Extract Queries*

Once the knowledge graph has been created (as a *Turtle* file), custom *SPARQL* queries are written to be executed against the data, one for each business competency question. The *SPARQL* queries are invoked from Python modules and the query results written to *CSV* files.

D. *Validate the data*

A system test *SPARQL* query was created to extract all data from the knowledge graph and validate the contents against the original input, to ensure no data has been lost or corrupted.

3.1.3 Self Service Analytics

A. *Extract data and load to BI*

The project is not extracting data to a Business Information tool.

B. *Answer Question*

The output of the custom extract queries is saved as *CSV* files and this should answer each of the business *competency questions* in turn.

3.2 Iterations

The Agile development methodology (Dingsøyr, 2012) has been much researched and used in general software development to deliver working software earlier and more frequently, allowing more timely evaluation and feedback from domain experts. It has also been applied in ontology engineering with the development of knowledge specific methods (Peroni, S. et al., 2017)) and has value in this project.

Applying Agile principles, such as daily meetings with the active involvement of stakeholders, were not possible but others ideas were used. Delivering a suite of working software at regular intervals, from source data to knowledge graph and query results, and responding quickly to changing needs or

requirements was applied to the project iterations. The revised plan estimated that 10 build iterations would be possible and these were all successfully completed (Table 1).

3.3 Project Management and Communication

All project communication between student and supervisors used online tools, that is, Microsoft Teams, Google Drive and email. Because of the lack of face-to-face contact, documenting project progress and sharing information efficiently with the supervisors was particularly important.

There were scheduled video conference meetings every 2 weeks with both supervisors and sharing of data on a project Google Drive[36]: meeting agendas and minutes, iteration descriptions and results and other useful project documents. All code and data were shared via the project's GitHub page[37].

The fortnightly supervisory meetings would review the completed iteration and discuss and agree the deliverables to be produced in the following 2 weeks. Planned deliverables would either be an increase in data scope or functionality of the Ontology and Knowledge Graph or the application of tools or methods from recent literature and an evaluation of whether they proved useful and would be carried forward in the software.

Details of the work completed in each iteration was shared in a folder on the project Google Drive. Each iteration folder included a graphical overview of the software completed, output results and other relevant information.

I presented an overview of the project[38] to my co-supervisor Valentina Carapella (Perspectum) and also kept a personal daily project diary of hours worked by task type, as well as descriptions of tasks completed, problems encountered and general reflections on the project.

On completion of the work, and with the co-supervisors, I took part in an hour-long presentation of the results of the project to over 30 Perspectum staff.[13]

3.4 Testing and Validation

3.4.1 Ontology

The completed ontology is validated for correctness using Protégé which will tell us whether the ontology has any unsatisfiable or inconsistent classes.

Another useful measure of quality assurance for the ontology can be achieved by using a basic checklist (Hitzler, 2021b). This includes validating that the ontology (a) uses disjointedness where appropriate (b) considering the nature of role characteristics (transitive, symmetric, etc) (c) avoiding

[36] Judith - Project - Google Drive (accessed 15/10/2021)
[37] https://github.com/JudithGrieves/City-MSc-Project (accessed 15/10/2021)
[38] Overview so far.pptx - Google Slides (accessed 2/11/2021)

Table 1. Description of iteration contents

Iteration	Start Date	Iteration Work Description
1	28 June 2021	• Create single draft ontology from supplied data • Define a mapping from input data to ontology entities • Code Python modules to create the knowledge graph from test CSV using the mapping • Code Python module to extract triples to CSV result set using *SPARQL* queries to answer competency questions.
2	5 July 2021	• Create PAYG Knowledge Report detailing high level requirements and details of data attributes • Create an Iteration Product Catalogue to document the changes proposed and completed in each iteration. • Split the ontology into Metric and Scanner sub-ontologies • Add link to QUDT unit of measure ontology
3	19 July 2021	• Evaluate RDF Mapping Language (RML) • Create an RML mapping file to map source data to the knowledge graph. • Code a Python module to execute the RMLmapper[27] program and create the knowledge graph.
4	2 August 2021	• Add OWL2 inferencing to the *SPARQL* query modules • Miscellaneous tweaks and bug fixes
5	16 August 2021	• Assessed the ontology and data so far against the *FAIR* principles. Document of results included (Appendix D). • Added concept and relationship descriptions from the Knowledge Report as semantic data in the ontology. • Add isDefinedby ontology annotations
6	30 August 2021	• Planned week of holiday • Literature review write-up • Create project products presentation slides and present to Valentina Carapella of Perspectum.
7	13 Sept 2021	• Unplanned week's break • Evaluate OTTR • Literature review writing
8	27 Sept 2021	• Draft report writing
9	11 Oct 2021	• Create Liver sub-ontology • Enhance all ontology annotations with skos:altLabel, Semantic Type and Group
10	25 Oct 2021	• Add external ontology links • Review and validate Liver sub-ontology with domain expert

specific *domains* and *range*s (d) distinguish correctly between parts and subclasses and (e) ensuring the correct direction of roles/relationships.

3.4.2 Knowledge graph

Whilst the *competency questions* evaluate whether the knowledge graph satisfies the business requirements, a system test was designed to ensure the mapping is correctly executed, that is, that all input data is fully and correctly represented in the output data. A *SPARQL* query was written to extract all data from the knowledge graph, output in the format of the input data, and a cell by cell, automated comparison of its results performed. In this way, we ensure that no data is lost in the transformation and any change to the code and data products can be quickly validated by re-running this test.

3.4.3 Implementation and User Documentation

The code modules, ontology and input data files are stored in a GitHub repository[37]. This includes basic User Documentation[39] to explain the contents of the repository and give directions on running the software.

3.5 Evaluations – tools and standards

Tool and method evaluations were carried out within iterations. After researching the background and user documentation of a tool, a quick usability test was carried out using sandbox data to ensure all technical requirements were met and the tool could be run successfully. When this initial test was successful the tool was implemented on the smaller of the project sub-ontologies (Scanner) and again, when successful, full implementation was made across the entire data scope.

A standard evaluation document was used, as far as possible, for each of the tools. This comprised an overview of the tool, description of the evaluation methods, evaluation criteria and results and, finally, the conclusions. Tools were evaluated for ease of use, documentation, debugging, etc.

[39] https://github.com/JudithGrieves/City-MSc-Project#user-guide (accessed 15/10/2021)

4 Results

This section describes the products created in the project: the Knowledge Report of requirements and data analysis, the ontology, the transform and load software and resulting knowledge graph, output queries and evaluation results.

4.1 PAYG Methodology - Knowledge Report

Business Questions	Answers
Why: Why do we need to answer these questions? What is the motivation?	Perspectum scientists wish to analyse previously gathered patient imaging and study data for various reasons. They may need to analyse this data to, for example, draw new biomedical conclusions from previous studies, to identify sub-cohorts of patient volunteers or examine trends in imaging data over time. At each stage of gathering and recording imaging study data, there are attributes of how each metric is obtained, stored or accepted. The study Metrics need to be accountable, reliable and reproducible and the data analyses should include attributes defining these factors.
Who: Who produces the data? Who will consume the data? Who is involved?	Patient and imaging metrics are gathered by the medical side of Perspectum. They conduct patient questionaires and assessments, perform MRI scans and record data on imaging success and quality, processing from patient assessment to storage of analysis output. Data Scientists perform additional computations on the output of this analysis to create validated metrics for each participant, producing final datasets on which statistics and analyses can be run. Biomedical Scientists can then use the final datasets to answer their questions.
How: How is this the business question answered today, if at all?	Currently Perspectum Data Scientists are the only people with the knowledge to manipulate and analyse the output of the medical assessments and must regularly spend time combining datasets, creating queries and producing results to answer questions from other areas of the company. This involves valuable time and resource and repetition of effort.

Name	Definition	ID (URI)	Source column	Source File / sheet
Patient	A human participant in the Perspectum long COVID study. For the purposes of this ontology, the patient is anonymously identified by a patient identifier, Pnnn, where n is a unique number.	URI_Patient_%patient id%, eg. URI_Patient_P275	Patient_ID	'TabularData'
ScanVisit	A visit made by a patient for the purpose of taking part in an MRI scan to gather metrics for the study data. A scan visit is defined as a single visit for a single patient and allows identification of multiple values of a single metric type for a single patient. A visit is currently identified by a sequential number currently given in the test data.	URI_Visit_%patient_id%_%visitID%, eg.URI_Visit_P005_2	Scan visit	'TabularData'
QuantitativeMetric	A quantitative metric gathered for the study. The metric may either be scan or patient related. Example metrics are Patient Age or Liver cT1.	see sub-types, LivercT1, etc	liver_cT1, liver_PDFF, liver_T2star	'TabularData'
LivercT1	Corrected T1 is an MRI derived liver biomarker providing a metric relating to inflammation and fibrosis. T1-relaxation time (measured in milliseconds) is a fundamental parameter in MRI relating to the interaction and energy exchange between the excited hydrogen atoms (usually in water) and the surrounding tissue structure. A sub class of Quantitative Metric. T1 measurements differ depending on magnetic field strength and MR manufacturer.	URI_LivercT1_%metricID%, eg. URI_liver_cT1_1000	liver_cT1	'TabularData'

Name	Definition	ID/URI	Concept	Source column
PatientSex	The biological sex of a patient, valid values 'Male'/'Female".	PatientSex	Patient	Sex
FieldStrengthValue	The field strength of the magnet used in an MRI Scanner, measured in teslas (T) Hospitals routinely use machines with field strengths of 1.5 T or 3 T, but ultra-high-field scanners are on the rise.	FieldStrengthValue	ScannerFieldStrength	Scanner Field Strength
FieldStrengthUnit	The unit of measure of the scanner field strength. Usually tesla (T)	FieldStrengthUnit	ScannerFieldStrength	Unit
MetricValue	The numeric value of a Quantitative Metric.	MetricValue	Quantitative Metric	liver_cT1, liver_PDFF, liver_T2star

Name	Definition	id (URI)	From Concept	Source Column (from)
isAttendedBy	The attendance at a scan visit of a patient in the study. A scan visit may be attended by many patients.	psp:isAttendedBy	ScanVisit	Scan visit
attendsVisit	The attendance of a patient at a scan visit. A patient may attend many visits.	psp:attendsVisit	Patient	Patient_ID
hasPatientMetric	to show a Patient has a particular Quantitative Metric recorded	psp:hasPatientMetric	Patient	Patient_ID
isMetricForPatient	to show a Quantitative Metric is recorded for a particular Patient.	psp:isMetricForPatient	Quantitative Metric	liver_cT1, liver_PDFF, liver_T2star
isMetricForVisit	to show a Quantitative Metric is recorded at a particular Scan Visit.	psp:isMetricForVisit	Quantitative Metric	liver_cT1, liver_PDFF, liver_T2star
hasVisitMetric	to show a scan visit has resulted in a particular Quantitative Metric	psp:hasVisitMetric	ScanVisit	Scan visit
usedInVisit	to show an MRI Scanner Model is used in a particular Scan Visit.	psp:usedInVisit	MRIScannerModel	Scanner
usesScannerModel	to show a Scan Visit used a particular MRI Scanner Model	psp:usesScannerModel	ScanVisit	Scan visit

Figure 8. Knowledge Report examples. (a) Requirements (b) Concepts (c) Attributes (d) Relationships

The *PAYG*-defined Knowledge Report was created with business questions and answers, data concepts, attributes and relationships. The report comprises 6 sheets of data (Appendix C). Figure 8 shows examples of the content.

These data descriptions were subsequently added to the ontology as semantic meta-data in iteration 5.

4.2 Iterative Development

Each iteration output comprised test input data, an updated ontology in *Turtle* format, custom coded Python modules, a resulting knowledge graph, SPARQL queries and output data. A graphical overview of the content of each iteration and ontology diagram were included in the shared project Google Drive, and where relevant, an evaluation report of a tool or method.

The most significant code enhancements were made in iterations 1-3 and described in the following sections.

4.2.1 Iteration 1

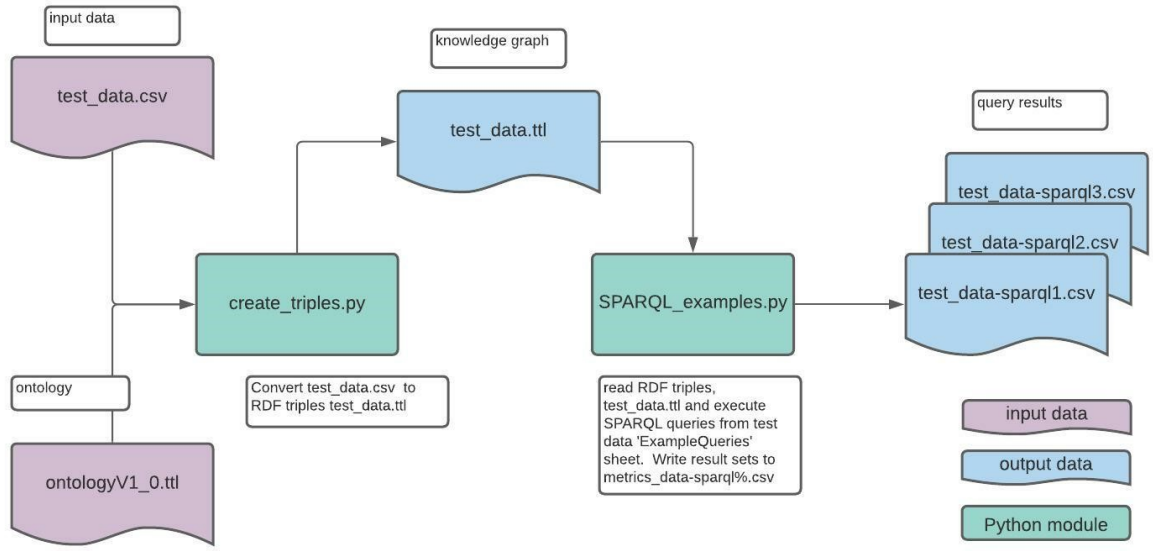


Figure 9. Overview of code created in Iteration 1

Iteration 1 (Figure 9) created the first version of the ontology based on information given with the input data and example expected triples. All data attributes were modelled in a single ontology and there were no references to any external vocabularies or identifiers. The transformation and load process was executed solely by custom coded Python modules. The Python modules were data driven as far as possible, that is, the knowledge graph was created based upon a set of rules declared in a Python list structure in the code.

4.2.2 Iteration 2

Iteration 2 (Figure 10) split the ontology into Metric and Scanner subsets, linked by the *URI* of the

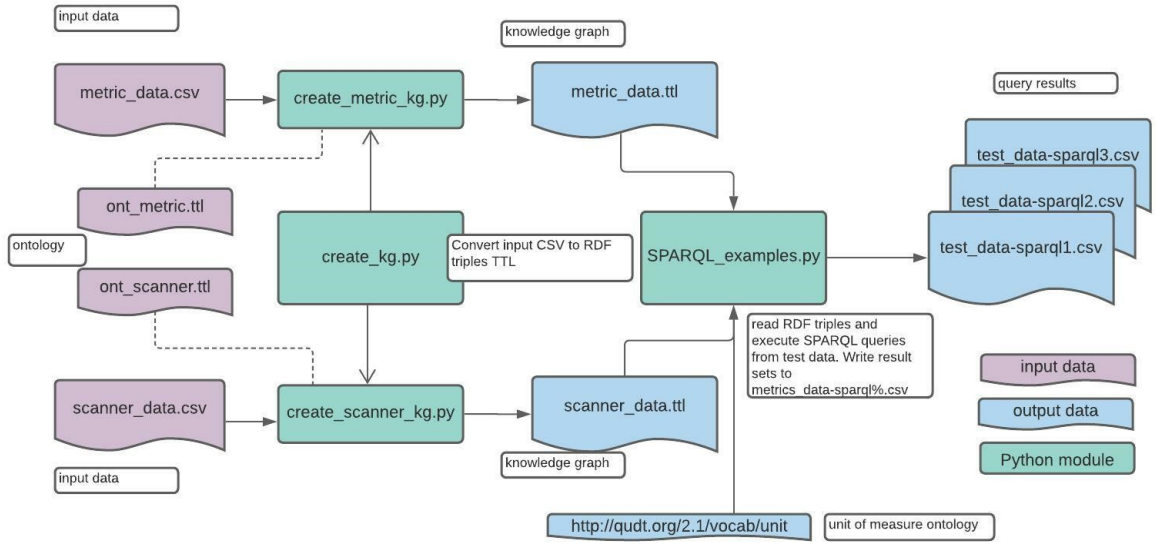


Figure 10. Overview of code created in Iteration 2

Scanner Model concept/class, and incorporated *URI* links to the QUDT[40] unit of measure ontology and vocabulary. The transformation and load modules were enhanced to abstract the knowledge graph creation for any given input file and mapping structure. This version of the software created two knowledge graph files, Metric and Scanner. The *SPARQL* queries were amended to read both knowledge graphs and to import the QUDT ontology to access unit of measure descriptions etc.

4.2.3 Iteration 3

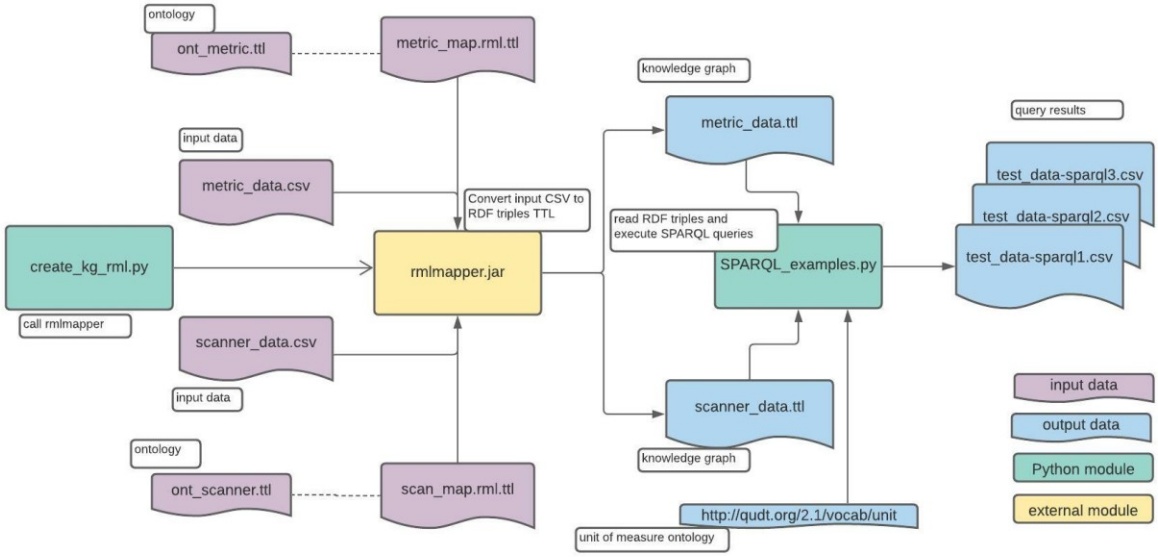


Figure 11. Overview of code created in Iteration 3

[40] http://qudt.org/ (accessed 5/10/2021)

Iteration 3 evaluated the RDF Mapping Language (RML) and it proved to be a useful tool to be used in the software. This involved the creation of RML mapping files for Metric and Scanner sub-ontologies and a new Python module to execute the open source RMLmapper.jar[29] to translate the input data files into knowledge graphs (Figure 11). The previously created Python transformation and load code was deprecated.

4.2.4 Evolution of the Ontology

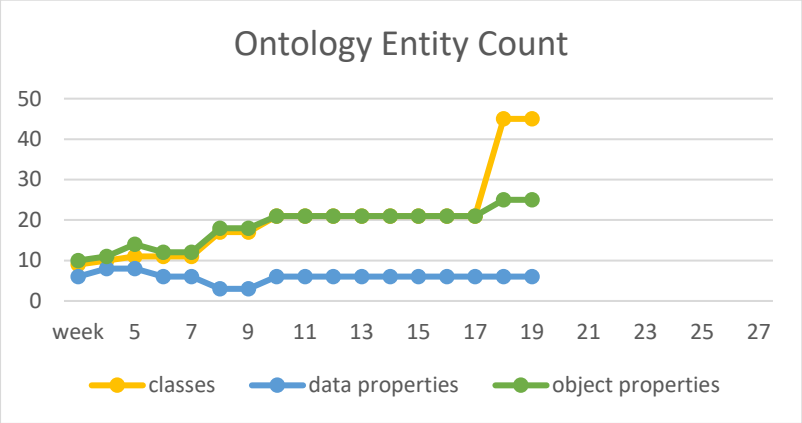


Figure 12. Ontology Entity Counts by week.

Metrics were recorded of the number of entities in the ontology (Figure 12) to chart its evolution over the project iterations. This shows the gradual increase in classes during initial development, followed by a plateau when the ontology stabilised and other aspects of the project, such as tool evaluation, were prioritised.

Another significant change was the decrease in data properties around week 7 – this reflects a move to convert data properties into new classes linked by *object properties* for increased flexibility of the model.

Week 17-18 saw a large increase when the liver context ontology was created to create a wider context for the data gathered.

4.3 Ontology

The final ontology has 45 classes, 25 *object properties* and 6 *data properties* (Figure 13). It comprises three sub-ontologies: Liver, shown as a Protégé hierarchy (Figure 14), Metric and Scanner, shown as graph diagrams (Figure 15 & Figure 16). This meets a stated Perspectum requirement for the ontologies to be broken down into components so that elements could be worked on independently and in parallel where necessary and also accords with the modular ontology paradigm.

	metric	scanner	Liver	TOTAL
Axiom	236	139	155	530
Logical axiom count	65	25	36	126
Declaration axioms count	40	23	31	94
Class count	16	7	22	45
Object property count	13	8	4	25
Data property count	4	2	0	6
Individual count	2	0	0	2
Annotation Property count	10	11	10	31

Figure 13. Ontology Metrics (Protege)

As with any knowledge modelling exercise, there were many design decisions to be made regarding the granularity and perspectives of the ontology. In modelling this domain, a balance was struck

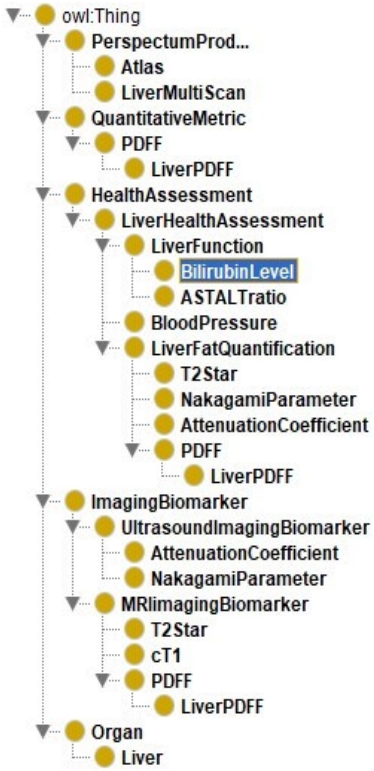


Figure 14. Sub-ontology 'Liver' (Protégé hierarchy)

between the detail of the domain knowledge and the data available to populate a knowledge graph. For example, some biological ontologies contain a great deal of subclass (subsumption) detail in the description of the anatomy of the liver but this was not deemed relevant for this project and therefore omitted.

The liver was defined as a subclass of 'organ' which provides interoperability to many existing ontologies and will allow the future addition of attributes for other organs relevant to Perspectum's work and was for this reason included.

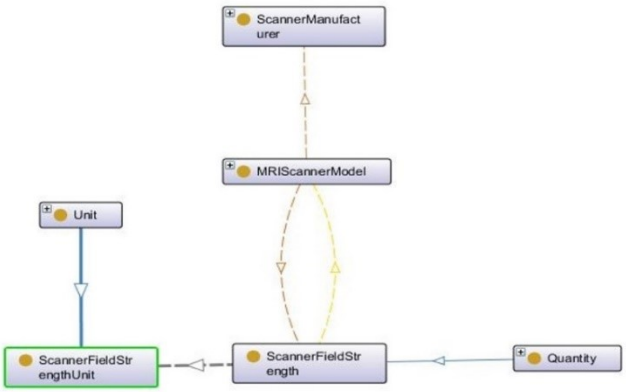


Figure 15. sub-ontology 'Scanner'

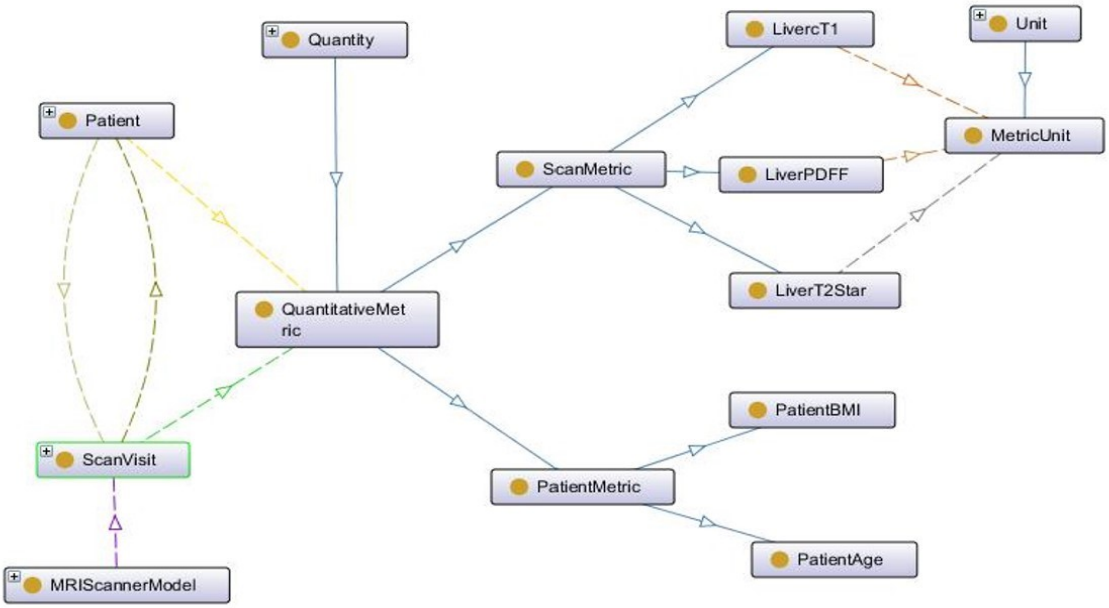


Figure 16. Sub-ontology 'Metric' linked to 'Scanner' by MRIScannerModel

The sub-ontologies are also available in the GitHub repository in OWL, XML and Turtle format[41].

Each ontology entity has associated meta-data, annotations of comments, labels and isDefinedby, populated from the information derived in the Knowledge Capture phase and documented in the knowledge report (Figure 17).

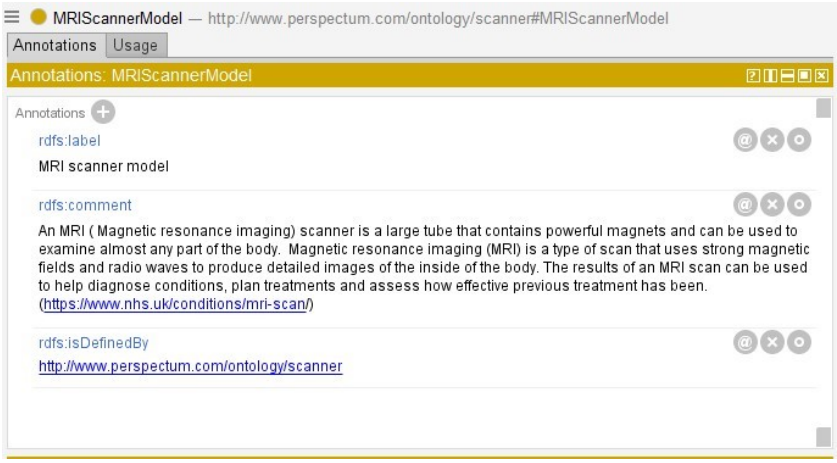


Figure 17. Protégé class annotations

Interoperability of the new ontology is provided by the inclusion of *URI* links to existing ontologies:

- QUDT: reuse of the URI in the project ontology
- CMR-QA (Figure 18)
- ONLIRA reuse of the URI in the project ontology

Figure 18. Related classes in CMR-QA ontology (Carapella et al, 2018)

4.4 Data Mapping/Migration

The transformation and load process creates the knowledge graph from the input tabular data. It consists of the definition of the data mapping and execution of the mapping to create the knowledge graph.

The mapping is a set of RML files to define the relationship between the source data and the ontology structure which was determined in the data analysis phase. Any future additions or changes to source data or mapping can simply be reflected in this file.

[41] City-MSc-Project/ontology at main · JudithGrieves/City-MSc-Project (github.com)

The RML mapping was manually created as a text file, containing details of the source data and a specification of the knowledge graph triples to be created. The triples' specification defines the subject, predicate and object maps with references to the tabular data columns from which to source the data. Figure 19 shows an example of the mapping, to create instances of the class 'MRIScannerModel' from 'scanner_model' source data column, and add related annotation triples of type 'rdfs:label 'and 'rdfs:isDefinedBy'.

A Python wrapper module was written to execute the transformation, using the RMLmapper[29] tool, and create the knowledge graph. RMLmapper execution requires the name of the mapping file and the output file to which the knowledge graph is to be written.

```
:scanner
  rr:subjectMap [
    rr:template "http://www.perspectum.com/resources/scanner/URI_MRIScannerModel_{scanner_model}";
    rr:class oscn:MRIScannerModel;
    rr:class owl:NamedIndividual
  ];
  rr:predicateObjectMap [
    rr:predicate rdfs:label;
    rr:objectMap [
      rml:reference "scanner_model";
      rr:datatype xsd:string
      ]
  ];
  rr:predicateObjectMap [
    rr:predicate rdfs:isDefinedBy;
    rr:objectMap [
      rr:template "http://www.perspectum.com/resources/scanner/"
    ]
  ].
```

Figure 19. RML mapping file example: definitions of triples for the Scanner class.

4.5 Knowledge Graph

The resulting knowledge graph is created from the input data via the RMLmapper[29] process and holds all instances of the data loaded. Figure 20 shows the creation of a section of the knowledge graph for patient visit P112/1 and related quantitative metrics. A row of tabular data results in a set of triples, based on the mapping, and the connected triples make up a section of the knowledge graph, visualised here using GraphDB[42].

The load process creates the knowledge graph as a *Turtle* format file. Figure 21 shows RDF triples for two instances of Liver cT1. The triples in this example belong to the class 'LivercT1', have attributes of Unit and Value and annotations of 'rdfs:label' and 'rdfs:isDefinedBy'. Relationship predicates 'isMetricForPatient' and 'isMetricForVisit' link to their respective Patient and Visit URIs.

[42] GraphDB™ - Ontotext (accessed: 3/12/2021)

Patient_ID	Scan visit	Scanner	liver_cT1	liver_T2star	liver_PDFF	Test_liver_volume	Gender	BMI	Age
P111	1	Siemens_ModelA	616.386597	26.6443853	4.194250647	1346.054683	F	32	35
P112	2	Siemens_ModelA	637.213511	13.323743	not quantifiable	1425.882797	F	18	45
P112	1	Siemens_ModelA	708.320852	17.379776	2.020770602	1317.326259	F	18	45
P113	2	Siemens_ModelA	571.007532	43.8729964	1.650659613	1263.545902	M	38	45
P113	1	Siemens_ModelA	842.313105	27.6011813	not quantifiable	1636.08133	M	38	45
P114	2	Siemens_ModelA	694.223875	36.0009418	3.723479303	1150.475305	M	22	19
P114	1	Siemens_ModelA	672.152694	38.2386718	6.043647126	1539.549337	M	22	19
P115	2	Siemens_ModelA	835.039083	32.6172275	not quantifiable	1417.152474	M	19	50
P115	1	Siemens_ModelA	692.34958	21.0706393	0.53512121	1588.449453	M	19	50

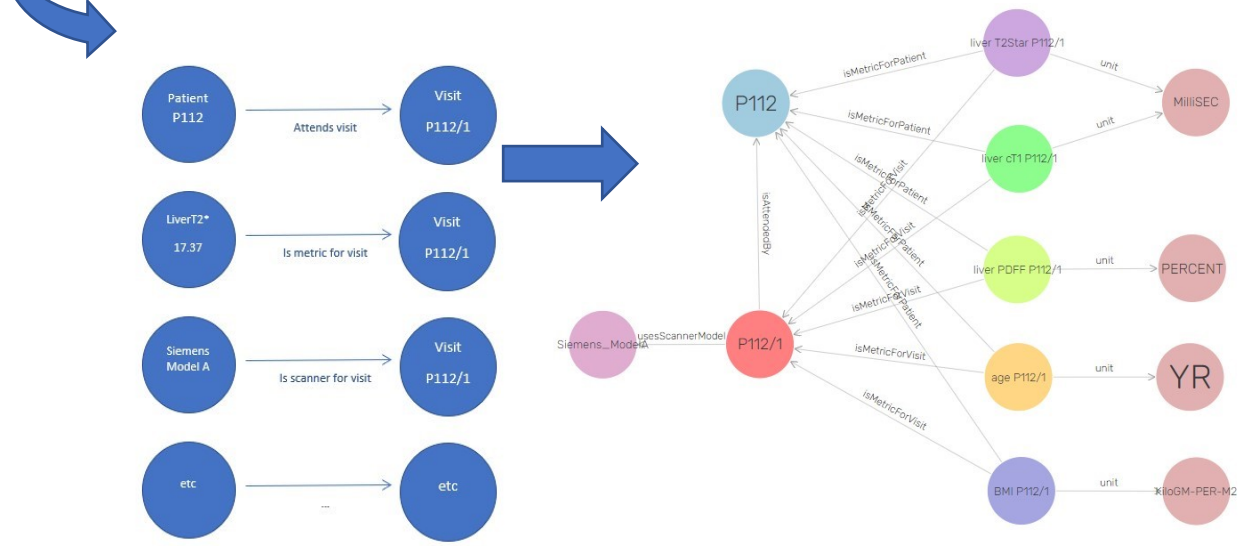


Figure 20. Transformation from tabular input data to knowledge graph for patient visit P112/1

```
met:URI_liver_cT1_P005_2 a omet:LivercT1, owl:NamedIndividual;
  qudt:unit unit:MilliSEC;
  qudt:value "799"^^xsd:float;
  omet:isMetricForPatient met:URI_Patient_P005;
  omet:isMetricForVisit met:URI_Visit_P005_2;
  rdfs:isDefinedBy <http://www.perspectum.com/resources/metric/>;
  rdfs:label "liver cT1 P005/2" .

met:URI_liver_cT1_P1000_2 a omet:LivercT1, owl:NamedIndividual;
  qudt:unit unit:MilliSEC;
  qudt:value "873"^^xsd:float;
  omet:isMetricForPatient met:URI_Patient_P1000;
  omet:isMetricForVisit met:URI_Visit_P1000_2;
  rdfs:isDefinedBy <http://www.perspectum.com/resources/metric/>;
  rdfs:label "liver cT1 P1000/2" .

met:URI_liver_cT1_P1001_2 a omet:LivercT1, owl:NamedIndividual;
  qudt:unit unit:MilliSEC;
```

Figure 21. Section of the knowledge graph in Turtle file format

4.6 Testing and Validation

4.6.1 Ontology Validation

The ontology is validated in Protégé for correctness and no logical inconsistencies are reported by the Reasoner.

4.6.2 Competency Questions

There were 3 business *competency questions* posed to evaluate this first sub-set of data used in the project. The questions are to show:

1. all Females with age above 40 and BMI above 25 that have liver cT1 above 800 ms
2. all Siemens 1.5 Tesla visits (patients scans) where PDFF is below 5%
3. all cases where cT1 is above 800 ms but PDFF is below 10%

The questions were posed by SPARQL queries executed within Python code (Figure 22) and provide the results in individual *CSV* files (Appendix I) that were compared with expected results generated from a manual filtering of the input data. The test result files show the expected and actual results with an automated comparison grid to highlight differences. In the example shown (Figure 23) an error is flagged for differences in the Visit column data, which on inspection is the results of a formatting difference where the Visit in the knowledge graph is identified by a concatenation of patient & visit.

```
cases where cT1 is above 800 ms but PDFF is below 10%
'''
qres = g.query(
"""SELECT DISTINCT  ?visit_label ?patient_label ?age ?bmi
                    ?livercT1 ?liverPDFF
        WHERE
            {?visit a ?ScanVisit .
            ?visit rdfs:label ?visit_label .
            ?visit omet:isAttendedBy ?patient .
            ?patient a ?Patient .
            ?patient rdfs:label ?patient_label .
            ?patient omet:hasPatientSex ?sex .
            {?PatientAge omet:isMetricForPatient  ?patient  .
             ?PatientAge omet:isMetricForVisit  ?visit  .
            ?PatientAge qudt:value ?age .
            ?PatientAge a omet:PatientAge  .}
            {?PatientBMI omet:isMetricForPatient  ?patient  .
             ?PatientBMI omet:isMetricForVisit  ?visit  .
            ?PatientBMI qudt:value ?bmi .
            ?PatientBMI a omet:PatientBMI  .}
            {?metric_PDFF omet:isMetricForPatient  ?patient  .
             ?metric_PDFF omet:isMetricForVisit  ?visit  .
            ?metric_PDFF qudt:value ?liverPDFF .
            ?metric_PDFF a omet:LiverPDFF  .}
            FILTER(?liverPDFF < 10 )
            {?metric_cT1 omet:isMetricForPatient  ?patient  .
             ?metric_cT1 omet:isMetricForVisit  ?visit  .
            ?metric_cT1 qudt:value ?livercT1 .
            ?metric_cT1 a omet:LivercT1  .}
            FILTER(?livercT1 > 800 )
            }
ORDER BY ASC(?patient_label)  """)
```

Figure 22. SPARQL to answer competency question 3

Compentency Question 3: cT1 > 800 ms ; PDFF < 10%																	
Expected Results							**Actual Results**						**Test Comparison**					
Patient_ID	Scan visit	Age	BMI	liver_cT1	liver_PDFF		Patient	Visit	Age	BMI	livercT1	liverPDFF	Patient	Visit	Age	BMI	livercT1	liverPDFF
P974	1	39	29.5	801	5		P974	P974/1	39	29.5	801	5	match	mismatch	match	match	match	match
P975	1	40	31	810	4.3		P975	P975/1	40	31	810	4.3	match	mismatch	match	match	match	match
P976	1	41	32.5	819	3.6		P976	P976/1	41	32.5	819	3.6	match	mismatch	match	match	match	match
P977	1	42	34	828	2.9		P977	P977/1	42	34	828	2.9	match	mismatch	match	match	match	match
P979	2	44	37	846	4		P979	P979/2	44	37	846	4	match	mismatch	match	match	match	match
P981	2	46	40	864	5		P981	P981/2	46	40	864	5	match	mismatch	match	match	match	match

Figure 23. Example Competency Question test results

4.6.3 System Test Results

The system test results compare the input data to a query on the output data to ensure that all data has been successfully transformed and loaded to the knowledge graph. An example is shown for a small sample of test data (Figure 24) where, as previously, the format of the Visit column data is different and shows as a mismatch in the Test Comparison.

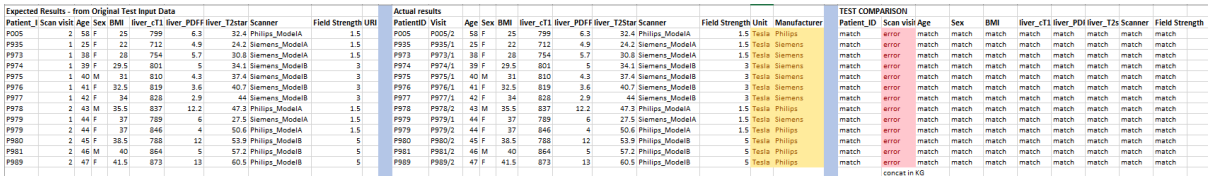


Figure 24. Example of System Test results

4.7 Evaluations – tools and standards

4.7.1 RDF Mapping Language (RML)

This evaluation of the RDF Mapping Language, RML mapper and related tools found that RML is an easy-to-understand method of mapping and generating RDF triples from *CSV* files. Of all the related tools aimed at simplifying the creation of the mappings, the creation of native RML in a text editor was found to be the most timely and straightforward means of creating a knowledge graph. Since this evaluation, it has been used to generate the knowledge graph of all current in scope project data.

RML is a set of declarative mappings showing the relationship between a semi-structured input dataset, in this case *CSV*, and an output knowledge graph consisting of *RDF* triples. It is an extension of R2RML, the World Wide Web Consortium (*W3C*) Relational Database (RDB) to RDF mapping language.

RML mappings can be created in several ways: directly using a text editor, via a higher-level, 'human readable' language, YARRRML[43], or by using a GUI which subsequently generates the file. A simple ranking method was used in the evaluation to compare each method, covering documentation, simplicity of use, development time, debugging etc. The manual creation of native RML mappings was the preferred method across all evaluation criteria (Table 2).

Following a successful proof of concept using the data in the Scanner ontology, native RML was created for the larger Metrics subset in a short time and was used for the ongoing code base in subsequent iterations.

Table 2. RML Evaluation Results

Tool	Documentation	Suitability for non-coders	Simplicity of overall process	Time to develop	Error likelihood	Error debugging	Total	Overall ranking
Native RML mapper	1	3	1	1	2	1	9	1
YARRRML	3	2	2	2	2	1	12	2
RMLEditor GUI	2	1	3	3	1	3	13	3

The success of the native RML mapping is a combination of simplicity and useability. It has a single mapping file and a single RDF creation process, with the mapping language being human-readable for use as documentation or in error correction. The RML mapping generated from both the YARRRML parser and the RMLEditor GUI were more difficult to read and debug. However, manually written RML does allow the possibility of syntax errors and any necessary debugging which the GUI does not.

[43] https://rml.io/yarrrml/ (accessed: 28/7/2021)

Creating the RML mappings directly using an editor was relatively straightforward, after an initial learning curve, and speedier than using the GUI. The process of generating the knowledge graph was a single-step process of running the RMLmapper versus a need for pre-parsing of the YARRRML file or export from the GUI.

YARRRML and RML native mappings both require knowledge of their syntax, though YARRRML is less verbose and therefore potentially more accessible.

The RML mapper error output is verbose but gives file line numbers for detected errors. Because the user enters native RML mapping rules and therefore becomes familiar with the syntax, debugging of errors was found to be reasonable.

Columns must have the same names as the mapping though can be in different order.

In this evaluation, *RML* has been used to create a new knowledge graph and an extension to this work would be to append new source data by merging into the existing graph.

4.7.2 FAIR Principles

The final ontology and knowledge graph were evaluated against the *FAIR* principles using documented implementation considerations (Jacobsen, A. *et al.*, 2020b). This resulted in an assessment of the extent to which the project ontology and data met each of the 15 *FAIR* principles for Findability, Accessibility, Interoperability and Reusability (Figure 25).

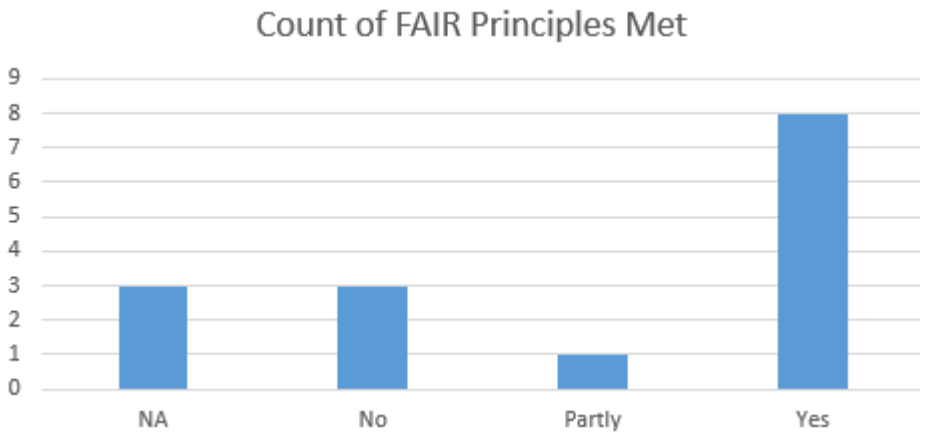


Figure 25. Results of FAIR Principles Evaluation

Eight of the 15 principles were deemed to be met and one partly. Of the others, 3 were not applicable: those dealing with licensing and web availability which are not and will not be applicable to data not intended for public consumption. Three principles not met were: inclusion of detailed data provenance, meeting relevant community standards and registered in a searchable resource; these principles were not met, due to unavailable data or lack of project time, but could be added if this work is extended in future.

To be findable, the project ontology contains unique identifiers, comprehensive annotations in the form of labels and comments as well as meta-data including the identifier of the data it describes. However, the data are not registered and indexed in a searchable resource as the intention is not currently to make this data public.

To allow Accessibility, the data use the RDF standards and HTTP protocols although there are no authorisation and authentication procedures deemed necessary at this stage. The data is *interoperable* by use of *W3C* recommended language OWL and qualified references to a number of other ontologies (QUDT, CMR-QA and onLIRA) all of which follow *FAIR* principles.

4.8 User Documentation

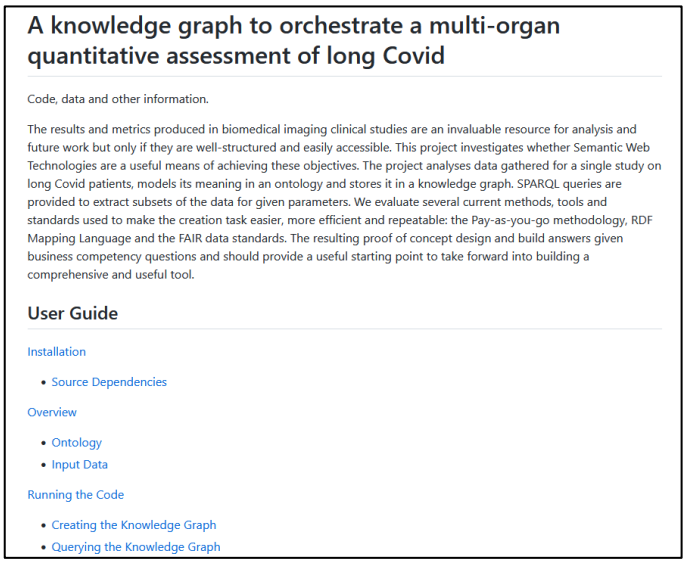


Figure 26. User Documentation on GitHub repository.

User documentation is included in the project GitHub repository[39]. It describes source dependencies, an overview of the contents of the repository and instructions for running the transformation and load code and querying the knowledge graph (Figure 26).

5 Discussion

The project successfully created a set of ontologies to model the problem domain and devised a relatively straightforward transformation and load process to generate a knowledge graph for any submitted test data. The knowledge graph was queried using SPARQL which was able to correctly answer a set of business *competency questions*.

This section considers these results in the context of the project objectives, original research questions and current literature.

5.1 Project Objectives and Research Questions

Objective A. To create an ontology, a systematic, formal, and unambiguous representation of the knowledge around imaging and other patient health measurements for the study data, characterising the profile of these values, that is, the data needed to unambiguously identify each one.

Research Question 1. Can ontologies provide a useful unified view of the complex use case presented by the Perspectum study data?

The Liver *MRI* measurements and related data was successfully modelled in 3 sub-ontologies: Liver, Metric and Scanner. The node-edge structure of an ontology allowed a granular representation of the features of the data required. Each imaging metric can be linked to the other attributes for which it is relevant, that is, the patient, visit and scanner for which it was recorded. Particularly important in future will be the addition of further provenance data to the study metrics, such as methods used in their collection, the stage of validation of an instance, etc. The granularity of the graph structure (versus other data management methods) will simplify these additions, without adversely affecting the existing model.

The graph structure allows any node to belong to multiple hierarchies, something that is not as easily represented in, eg. Relational modelling.

The ontology annotations allowed the inclusion of comments, text and references that provide detailed and useful knowledge to future users of the data, human and machine, and increase the findability of the data.

The ontology contained *URI*s identifying other ontologies and vocabulary and thus has a degree of interoperability, that is, can link to and access a body of external information.

Objective B. To create a knowledge graph as instances of these concepts from tabular input data.

Research Question 2. Can the data in this domain, that is, measurements related to liver imaging clinical studies, be put into context using a knowledge graph? Are the hierarchies and relationships at the knowledge graph level useful to profile this data?

The knowledge graph was shown to be an appropriate data structure in which to hold imaging measures and study data. The input tabular data was easily mapped to the ontology structures and transformed into triple stores that could be queried to correctly answer the *competency questions*.

The structure in the knowledge graph will allow further questioning; data can be queried for any of the hierarchies. Data can be extracted, for example, for a single patient, a single study, a particular *MRI* machine or magnetic field strength, allowing comparisons to be made and conclusions drawn.

This work takes a specific format of tabular input dataset and creates a knowledge graph. However, it has not so far considered how and with what frequency these inputs will be added to the triples. It is expected that a business implementation would consider: pre-processing steps to validate input data, the merging of new data into the graph, updates of existing URIs and generally ensure the smooth running of the process, accounting for all eventualities.

Objective C. To investigate and evaluate currently researched methods of creating this knowledge data that are robust, repeatable, and accurate.

Research Question 3. Can we (semi)automate and streamline the creation of the ontology and knowledge graph via templates and other recently researched methods?

Research Question 4. What features of ontologies and knowledge graphs recommended in current literature should be included to enhance the usefulness of the products?

The Pay-as-you-go methodology (Sequeda et al, 2019) provided a useful way to structure the initial analysis and provide documentation of the data in scope. The requirements questions of what, why, who, how, where and when were a targeted way of eliciting the overall business requirements in discussions with Perspectum. Answering the business questions documented the current issues experienced by Perspectum and the improvements required. The descriptions of classes, *object properties* and *data properties* gave a clear understanding of how the data should be modelled and encouraged questioning and validating the understanding of the data at a granular level. These descriptions also provided text to incorporate into semantic descriptions in the ontology to provide persistent understanding.

The custom Python coding to execute the mapping of source data to knowledge graph was initially straightforward. However, as the requirements of the knowledge graph were added and gradually became more detailed, the resulting code became more complex and time-consuming to extend.

Fortunately, the successful evaluation of RDF mapping language (RML) and RMLmapper provided a useful tool to declare and execute the mapping, avoiding complex coding in Python. Given the declarative mapping to link source data elements to ontology entities, it was a straightforward process to create the RDF knowledge graph.

As well as providing the mapping input to the RMLmapper, the mapping file serves as a reliable and up to date document of the relationship between source and target. RML can easily be amended or extended in line with changes to input data contents or ontology and knowledge graph structure and, as the RML mapping file is written in RDF, comments can be added to aid understanding.

Evaluation of the FAIR data principles provided useful validation and feedback, used to enhance the ontology and improve its adherence to these standards.

The project did not have time to apply modular ontology modelling so it remains an outstanding task to evaluate the success and advantages of this method of ontology creation.

The OTTR ontology had seemed a valuable way of abstracting patterns in the ontology and hence simplifying future enhancements but in practice proved less than straightforward to implement.

Objective D. To investigate current methods of accessing semantic data so that users can benefit from the querying capabilities of knowledge graphs

There was insufficient time during the project to comprehensively research and evaluate user-friendly query tools so data is extracted using *SPARQL* queries. These were custom coded for each of the *competency questions* and the results were used for an evaluation of the graph. SPARQL was able to answer all the posed questions and provide the results required for test and comparison extracts. Depending on the specific querying requirements on the data which emerge from wider use, a longer-term solution could be a blend of custom queries and the extraction of larger, more generic results that are then analysed in more user-friendly tools. Template-based query forms driven by the ontology or state of the art tools like OptiqueVQS[44] should be straightforward to implement.

Research Question 5. Is there a feasible alternative to the creation of a new Semantic data source, for example, the use of an ontology and mappings to access the data in a non-Semantic form or the use of alternative database technologies such as relational?

The project did not fully explore this research question.

[44] https://sws.ifi.uio.no/project/optique-vqs/ (accessed: 5/12/2021)

Given that the source data exists in disparate spreadsheets, it is not likely that ODBA mapping from source to ontology would be a viable option.

Although there was insufficient time to create a parallel repository using relational database technology and make a quantitative comparison, my previous experience in relational modelling and database design has identified several benefits of Semantic Web.

- The semantic knowledge that can be created, as an intrinsic part of an ontology in meta-data text and labels, is a valuable and persistent resource for future domain users, data analysts and software developers and the visualisation of the graph model is a user-friendly format for communication of domain knowledge. In relational database projects, data modelling and documentation, though frequently thorough, is technically separate to and independent of the database. This requires a systematic discipline to access and maintain in future developments to the data.
- The granularity of ontology classes and relationships should make extensions easier to incorporate without adversely affecting existing structures. In relational models and databases, new or changing knowledge and requirements often lead to changes that impact established entities or tables.
- The graph structure allows flexible querying over any set of data; relational databases often have to create multiple data marts based upon particular requirements.
- The linking to open-source external ontologies and knowledge graphs via URIs provides a link to vast existing knowledge and resources.

5.2 Current Literature

The literature consistently highlights the need for ontologies to strive for reuse of existing, well-used and respected ontologies. This project had to balance this ideal with the constraints of time available and the requirement for clarity for domain experts. Although there were 2 previous examples of imaging data ontologies found in the research, neither were available online. Hence, there was no existing ontology for the project domain to reuse or learn from. The ontology does however contain links to existing publicly available ontologies with some areas of overlap, QUDT, OnLIRA and CMR-QA, and hence provides a degree of interoperability whilst tailoring the structure and granularity to the specific data requirements of the Perspectum case.

The use of a foundational ontology is also specified as a desirable feature of new models. This requirement had to be balanced against development time and usability of the project end products as foundational ontologies can be abstract and difficult to appreciate and use without time and experience.

Current literature specifies a number of key requirements for ontology development. The *FAIR* data standards require ontologies and knowledge graphs to be Findable, Accessible, Interoperable and

Reusable. The project ontology and knowledge graph were evaluated and met or partly met 9 of the 12 applicable principles. The principles were a useful resource; with an element of rework being done after the first assessment to improve the evaluation.

6 Evaluation, Reflections, and Conclusions

6.1 Evaluation

6.1.1 Objectives

The project has been able to confidently meet the first 2 project objectives, that is, to model the domain data unambiguously in an ontology and create a knowledge graph of data in a corresponding structure. The ontology and knowledge graph were able to successfully answer the posed *competency questions*.

The third objective, of robust, repeatable and accurate methods, was partly met, with the use of RML as the transformation and load process to generate the knowledge graph and the application of the *FAIR* principles to evaluate and improve the ontology. Ideally, given more time, we would have also used modular ontology modelling to abstract common patterns and structures from the data.

The final objective was met but not as comprehensively as intended. The intention was to find a means of querying the data that did not require detailed knowledge of the data structures. More research is required to investigate this aspect more fully. SPARQL queries were created, and successfully returned correct results, but any further querying of the data will require skills in the query language and an understanding of the ontology structure. However, the use of SPARQL does allow the possibility of a hybrid solution which could extract generic grids of data that can then be queried separately using Business Intelligence tools.

Research question 5, on alternative modelling paradigms, was not fully explored but a number of advantages of ontologies and knowledge graphs (over relational models) were found for this particular domain. Ontologies were seen as superior repositories of knowledge and communication, allowed more flexible querying and ease of enhancement and addition.

On reflection, the number and scale of project objectives was quite ambitious given the project timescales. The Perspectum client is satisfied with the outcome of this proof of concept and hopes it will lead to a larger project to cover all objectives.

6.1.2 Literature Review

The literature review did not find any current publicly available ontologies covering this domain of imaging study data and metrics, providing some justification of the value of this project as a means of covering this gap in the research.

Some way through the project, while further researching related ontologies, I discovered a rich seam of imaging metric research papers previously unseen. This highlighted omissions in the initial search and how the literature review process can and should be iterative and continue, inspired by new findings, as the project progresses.

6.1.3 Methods and Planning

The revised project plan was a helpful roadmap and was followed reasonably closely and the iterations were a useful way of structuring the work and progress, reviewing regularly and learning lessons to improve the process. I would use it again and recommend to others the use of the Pay-as-you-go methodology.

The discipline of documenting the end of each 2-week iteration, ostensibly as a means of communicating progress to the project supervisors, was useful to order my thoughts and feed into what to tackle next. The iterative development was useful as each step was evaluated afterwards to learn lessons that could be applied to the next phase.

6.2 Reflections

One of the most important things when Modelling ontologies is close collaboration with the domain experts, to fully understand the data and validate the model. Because of the nature of this academic project the ideal amount of collaboration was not possible so the resulting ontology should be treated as a first draft to be thoroughly validated when used in practice.

The methods I used to control and document the project were useful. For a part time student, a daily task diary was helpful to track how many hours were being spent each week and on which tasks. The diary descriptions helped to recall unfinished tasks, understand what was not completed and what to avoid when restarting.

Data modelling and loading is a more qualitative task than most previous work in the MSc and finding quantitative measures in order to evaluate the work was therefore more difficult.

With more time to spend on the project I would have liked to successfully implement the OTTR tool and a method of abstracting patterns in the ontology. The identification of these patterns would give a greater understanding of similarities across the data and make expansion of the ontology more controlled and less prone to error in future.

A frequently quoted drawback of ontologies and knowledge graphs is the requirement for people with a deep understanding of the technologies to build and maintain them. This project has shown that the skillset suggested in the *PAYG* methodology (Sequeda, 2019) is sufficient. That is, a Knowledge Scientist should have broad technical skills, a background in data modelling and SQL, as well as good communication skills to interact with business users and produce readable documentation.

6.3 Conclusions

This project has shown that Semantic Web technologies are a useful means of understanding and holding the quantitative metrics from imaging clinical studies

An ontology is an appropriate method of modelling the knowledge and provides a user-friendly understanding and clarity. The knowledge graph based on the ontology and generated from tabular input data provides a means to extract results flexibly and can answer the required business *competency questions*.

The ontology model and transformation and load processes should be straightforward to enhance and extend. Any extension to the data scope requires a further iteration of the tasks identified: analysis of the new data, relationships and evaluating *competency questions*, modelling in the ontology, amendment of the RML mapping file and creation of additional SPARQL queries.

The project showed that the RDF mapping language (RML) is a powerful way to define a data mapping and generate a knowledge graph, simplifying the process and avoiding the need for custom coding. The *FAIR* principles were found to be helpful in validating the data structures and including some of the unmet principles of including data provenance and adhering to relevant community guidelines will be useful in future work.

It is recommended that Perspectum continue to explore and expand the scope of the work.

6.4 Future work

- This project has been carried out on small, manageable input datasets with a narrow scope. Addition of a broader range of data attributes from the problem domain should be straightforward and provide more value. Similarly, a useful future task would be to experiment with performance testing of the software for larger datasets.
- Currently a single input dataset creates a single knowledge graph. Enhancement of the software should be made, to allow repeated addition of new datasets and their incorporation into the single knowledge graph.
- A strategy to provide appropriate and user-friendly access to query the data should be formulated to provide significant business advantage from the data. This may simply involve SPARQL extracts tailored to specific requirements.
- More tools and methods research and evaluations, particularly in the use of ontology design patterns, would be useful and may rationalise development.

Glossary

Term	Description
biomarker	A 'biological marker', defined by the World Health Organisation as 'almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction'. Biomarkers can include simple body measurements from pulse and blood pressure through basic chemistries to complex results of tests on blood and other tissues. In this domain they mainly refer to the quantitative metrics obtained from MRI scans.
Competency questions	A business question defined as a test of the success of data structures and instances.
CSV	Comma Separated Variable, describing a format of file where the tabular values are delimited by a comma.
data properties	An element of data belonging to an ontology class and given a data type, eg. Boolean, string, numeric.
domain	In the context of an ontology object property (or relationship), the domain defines the source class of the relationship.
FAIR	Findable, Accessible, Interoperable and Reusable: a set of principles to aid semantic data modelling.
Findable	A FAIR principle defining how ontology elements should be held and described in a way that allows other agents to find them.
First-class entity	An item of data, in modelling a domain, that is important in it's own right, that is, has important relationships and attributes of its own and is not simply an attribute of another.
Foreign Key	In relational modelling and databases, a foreign key is a data attribute of an entity or table pointing to the primary key of another.
HTTP	Hypertext Transfer Protocol, the foundation of world wide web communication allowing requests and responses between client computers and web servers.
Interoperable	A feature of an ontology allowing it to be interconnected with other ontologies and knowledge graphs. Interoperable is one of the FAIR principles.
JSON	JavaScript Object Notation, a format for storing and sending data on the web.
MRI	Magnetic Resonance Image
object properties	A relationship defined in an ontology between 2 classes.
OWL	The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things.
PAYG	Pay-as-you-go methodology (Sequeda 2019). A series of project steps to be completed to model an ontology and build a knowledge graph for a subset of source data and business questions.
RadLex	RadLex (Radiology Lexicon) is a controlled vocabulary developed by the Radiological Society of North America (RSNA) for the purpose of providing a unified source of radiology terms that aims to reduce ambiguity in imaging related data and technology (Kundu, S. et al, 2009),
range	In the context of an ontology object property (or relationship), the range defines the target class of the relationship.
Resource Definition Language (RDF)	RDF is a standard means of data definition and interchange on the web, consisting of triples to define a graph data structure.
SNOMED CT	The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) covers a broad range of clinical terminology (SNOMED CT website, 2017) that aims to be the standard used in Electronic Health Records (HER) systems.

Term	Description
SPARQL	The language used to define queries on an RDF data source.
triples	The data structure underlying RDF, consisting of a subject, predicate and object to define the relationship between two nodes.
Turtle format	Terse RDF Triple Language is a file format and syntax to express RDF. Turtle is easily read and interpreted.
UMLS	The Unified Medical Language System (UMLS) (Bodenreider, 2004) was initiated by the U.S. National Library of Medicine in 1986 as a system for merging and mapping vocabulary from over 130 different sources to promote creation of more effective and inter-operable biomedical information systems and services.
Universal Resource Identifier (URI)	A unique identifier that each resource in the Sematic Web must have.
W3C	W3C is the World Wide Web Consortium, the international Semantic Web standards community.
XML	eXtensible Markup Language, a means of storing and sending web data, designed to be human and machine readable

References

Amdouni, E. and Gibaud, B. (2018) 'Imaging Biomarker Ontology (IBO): A Biomedical Ontology to Annotate and Share Imaging Biomarker Data', *Journal on Data Semantics*, 7(4), pp. 223–236. doi:10.1007/s13740-018-0093-3.

Arenas-Guerrero, J. *et al.* (2021) 'Knowledge Graph Construction with R2RML and RML: An ETL System-based Overview', p. 15.

Ashburner, M. *et al.* (2000) 'Gene Ontology: tool for the unification of biology', *Nature genetics*, 25(1), pp. 25–29. doi:10.1038/75556.

BERNERS-LEE, T. et al., (2001) 'THE SEMANTIC WEB', *Scientific American*, 284(5), pp. 34–43.

Berners-Lee, T. (2006) *Linked Data - Design Issues*, *Linked Data Principles*. Available at: https://www.w3.org/DesignIssues/LinkedData.html (accessed: 11 June 2021).

Bezerra, C. et al, (2013) 'Evaluating Ontologies with Competency Questions', in, pp. 284–285. doi:10.1109/WI-IAT.2013.199.

Buckler, A.J. *et al.* (2013) 'Quantitative imaging biomarker ontology (QIBO) for knowledge representation of biomedical imaging biomarkers', *Journal of Digital Imaging*, 26(4), pp. 630–641. doi:10.1007/s10278-013-9599-2.

Callahan, T. J. *et al.* (2020) 'Knowledge-Based Biomedical Data Science', *Annual review of biomedical data science*, 3, pp. 23–41. doi: 10.1146/annurev-biodatasci-010820-091627.

Carbon, S. et al., 2018, "The Gene Ontology Resource: 20 years and still GOing strong", *Nucleic acids research,* vol. 47, no. D1.

Carapella, V. *et al.* (2016) 'Towards the Semantic Enrichment of Free-Text Annotation of Image Quality Assessment for UK Biobank Cardiac Cine MRI Scans', *Deep Learning and Data Labeling for Medical Applications*, pp. 238–248. doi: 10.1007/978-3-319-46976-8_25.

Cogan, S. et al., (2020) Modular Ontology Modeling | www.semantic-web-journal.net. Available at: http://www.semantic-web-journal.net/content/modular-ontology-modeling/ (accessed: 15 June 2021).

Dimou, A. et al. (2013). Extending R2RML to a source-independent mapping language for RDF. *CEUR Workshop Proceedings*. 1035.

Dimou, A. *et al.* (2014) 'RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data', in. *CEUR Workshop Proceedings*.

Dimou, A. (2020) 'R2RML and RML Comparison for RDF Generation, their Rules Validation and Inconsistency Resolution', *arXiv:2005.06293 [cs]* [Preprint]. Available at: http://arxiv.org/abs/2005.06293 (accessed: 17 August 2021).

Dingsøyr, T. *et al.* (2012) 'A decade of agile methodologies: Towards explaining agile software development', *Journal of Systems and Software*, 85(6), pp. 1213–1221. doi: 10.1016/j.jss.2012.02.033.

FAIRsharing: QUDT; Quantities, Units, Dimensions and Types (2021); DOI: https://doi.org/10.25504/FAIRsharing.d3pqw7 ; (accessed: 7 August 2021)

Hitzler, P. (2021a) 'A review of the semantic web field', *Communications of the ACM*, 64(2), pp. 76–83. doi: 10.1145/3397512.

Hitzler P. et al. (2021b) *Foundations of Semantic Web Technologies*. pp307-334. Available at: http://web.a.ebscohost.com/ehost/ebookviewer/ebook/bmxlYmtfXzE3NjM1OTNfX0FO0?sid=dee7f62a-367c-491a-a386-aaa2a26fa04b@sdc-v-sessmgr03&vid=0&format=EB&rid=1 (accessed: 9 August 2021).

Hoehndorf, R et al. (2015) 'The role of ontologies in biological and biomedical research: a functional perspective', *Briefings in Bioinformatics*, 16(6), pp. 1069–1080. doi:10.1093/bib/bbv011.

Hogan, A. (2019) *The Semantic Web: Two Decades On | www.semantic-web-journal.net*, *Semantic Web Journal*. Available at: http://semantic-web-journal.net/content/semantic-web-two-decades-0# (accessed: 17 June 2021).

Hutton, C. *et al.* (2018) 'Validation of a standardized MRI method for liver fat and T2 quantification', *PloS one*. Edited by F. Bonino, 13(9), pp. e0204175–e0204175. doi:10.1371/journal.pone.0204175.

Jackson, R.C. et al. (2019), "ROBOT: A Tool for Automating Ontology Workflows", BMC bioinformatics, vol. 20, no. 1, pp. 407-407.

Jacobsen, A. *et al.* (2020a) 'A Generic Workflow for the Data FAIRification Process', *Data Intelligence*, 2(1–2), pp. 56–65. doi: 10.1162/dint_a_00028.

Jacobsen, A. *et al.* (2020b) 'FAIR Principles: Interpretations and Implementation Considerations', Data Intelligence, vol. 2, no. 1-2, pp. 10-29.

Jiménez-Ruiz, E. *et al.* (2015) 'BootOX: Practical Mapping of RDBs to OWL 2', in Arenas, M. et al. (eds) *The Semantic Web - ISWC 2015*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 113–132. doi:10.1007/978-3-319-25010-6_7.

Kharlamov, E. et al. (2015), "Ontology Based Access to Exploration Data at Statoil" in Springer International Publishing, Cham, pp. 93-112.

Kökciyan, N. *et al.* (2014) 'Semantic Description of Liver CT Images: An Ontological Approach', *IEEE Journal of Biomedical and Health Informatics*, 18(4), pp. 1363–1369. doi: 10.1109/JBHI.2014.2298880.

Kundu, S. *et al.* (2009) 'The IR Radlex Project: An Interventional Radiology Lexicon—A Collaborative Project of the Radiological Society of North America and the Society of Interventional Radiology', *Journal of Vascular and Interventional Radiology*, 20(4), pp. 433–435. doi: 10.1016/j.jvir.2008.10.022.

Marwede, D. et al., (2007) 'RadiO: A Prototype Application Ontology for Radiology Reporting Tasks', *AMIA Annual Symposium Proceedings*, 2007, pp. 513–517.

Kong, Y.M. *et al.* (2011) 'Toward an ontology-based framework for clinical research databases', *Journal of biomedical informatics*, 44(1), pp. 48–58. doi:10.1016/j.jbi.2010.05.001.

Messaoudi, R. et al. (2019a) 'Ontology-Based Approach for Liver Cancer Diagnosis and Treatment', Journal of Digital Imaging, 32(1), pp. 116–130. doi: 10.1007/s10278- 018-0115-6

Messaoudi, R. et al. (2019b) 'An Ontological Model for Analyzing Liver Cancer Medical Reports'. In Information Systems, edited by Marinos Themistocleous and Paulo Rupino da Cunha, 369–82. Lecture Notes in Business Information Processing. Cham: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-11395-7_29.

Messaoudi, R. et al., (2020); "Ontologies for Liver Diseases Representation: A Systematic Literature Review", Journal of digital imaging, vol. 33, no. 3, pp. 563-573.

Musen, M.A et al. (2012), "The National Center for Biomedical Ontology", Journal of the American Medical Informatics Association : JAMIA, vol. 19, no. 2, pp. 190-195.

Musen, M. (2015) 'The protégé project', AI Matters, 1, pp. 4–12. doi:10.1145/2757001.2757003.

Noy, N. et al. (2010) 'The ontology life cycle: Integrated tools for editing, publishing, peer review, and evolution of ontologies', AMIA Annual Symposium Proceedings, 2010, pp. 552–556.

Peroni, S. (2017) 'A Simplified Agile Methodology for Ontology Development', in Dragoni, M., Poveda-Villalón, M., and Jimenez-Ruiz, E. (eds) OWL: Experiences and Directions – Reasoner Evaluation. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 55–69. doi:10.1007/978-3-319-54627-8_5.

Roldán-García, M. del M. et al. (2018) 'Towards an ontology-driven clinical experience sharing ecosystem: Demonstration with liver cases', Expert systems with applications, 101, pp. 176–195.

Sansone, S.-A et al. (2019) FAIRsharing as a community approach to standards, repositories and policies. Nature biotechnology, 37, 358: 10.1038/s41587-019-0080-8. (accessed: Aug 17 2021)

Sequeda, J.F. et al. (2019), "A Pay-as-you-go Methodology to Design and Build Enterprise Knowledge Graphs from Relational Databases" in Springer International Publishing, Cham, pp. 526-545.

Shimizu, C. et al (2019) 'MODL: A Modular Ontology Design Library'. Available at: https://arxiv.org/abs/1904.05405.

Shimizu, C. et al. (2020), "The enslaved ontology: Peoples of the historic slave trade", Web semantics, vol. 63, pp. 100567.

Shimizu, C. et al (2021) *Modular Ontology Modeling | www.semantic-web-journal.net*. Available at: http://www.semantic-web-journal.net/content/modular-ontology-modeling (accessed: 15 June 2021).

Skjæveland, M.G. et al. (2018a), "Practical Ontology Pattern Instantiation, Discovery, and Maintenance with Reasonable Ontology Templates" in Springer International Publishing, Cham, pp. 477-494.

Skjæveland, M. et al. "Semantic Material Master Data Management at Aibel." International Semantic Web Conference (2018b).

Smedley, D. *et al.* (2021) '100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report', *The New England Journal of Medicine*, 385(20), pp. 1868–1880. doi:10.1056/NEJMoa2035790.

Smith, Barry, and Richard H. Scheuermann. (2011), 'Ontologies for Clinical and Translational Research: Introduction'. *Journal of Biomedical Informatics* 44, no. 1: 3–7. https://www.sciencedirect.com/science/article/pii/S1532046411000049?via%3Dihub.

The Gene Ontology Consortium *et al.* (2021) 'The Gene Ontology resource: enriching a GOld mine', *Nucleic Acids Research*, 49(D1), pp. D325–D334. doi:10.1093/nar/gkaa1113.

Tu, S. et al. (2009) 'OCRe: An Ontology of Clinical Research'.

Wahab, K. *et al.* (2019) 'Building a Biomedical Ontology for Chronic Liver Disease', in *2019 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, pp. 1–5. doi: 10.1109/CITS.2019.8862104.

Whetzel, P.L. *et al.* (2011) 'BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications', *Nucleic Acids Research*, 39(Web Server issue), pp. W541–W545. doi:10.1093/nar/gkr469.

Wilkinson, M. D. *et al.* (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3(1), p. 160018. doi: 10.1038/sdata.2016.18.

Xiao, G. *et al.* (2020) ‘The Virtual Knowledge Graph System Ontop: 19th International Semantic Web Conference on Demos and Industry Tracks: From Novel Ideas to Industrial Practice’, in.

Appendices

Appendix A RMPI Project Proposal for MSc in Data Science

Name: Judith Grieves

E-mail address: judith.grieves@city.ac.uk

Contact Phone number: 07979 915 095

Project Title: A knowledge graph to orchestrate a multi-organ quantitative assessment of long Covid

Supervisor: Ernesto Jiménez-Ruiz

A knowledge graph to orchestrate a multi-organ quantitative assessment of long-COVID

Introduction

This project will investigate the feasibility of using Semantic Web Technologies to load, store and access knowledge and metrics about multi-organ imaging of patients. Specifically, it will research the benefits and issues of using an ontology, to model the knowledge, and Research Definition Framework (RDF) data to store and access the results of a clinical study.

Semantic Web Technologies encompass the modelling of knowledge and data in a machine-accessible way that facilitates the interlinking and sharing of that data (Hitzler et al, 2010). It is already used in many scientific/bio-medical settings and has potential to be a useful technology to apply in this specific use case.

Perspectum (2021) is an Oxford-based medical imaging company, founded in 2012, specialising in non-invasive multi-organ imaging to assess patient health. The company has published a longitudinal, observational clinical study (Dennis et al, 2020) of patients, still experiencing symptoms 3 months post-infection, with acute SARS-CoV-2. Patients were assessed with health questionnaires, blood investigations and quantitative magnetic resonance imaging (MRI).

Motivation

Perspectum currently holds this study data in multiple tabular datasets and would like to formulate a more comprehensive approach to storing, querying, and publishing this data.

Objectives

The objectives of this project are:

- To create a systematic, formal, and unambiguous representation of the knowledge, around imaging and other patient health biomarkers for the study data, characterising a metric's profile, that is, the data needed to unambiguously identify each metric.
- To investigate and evaluate methods of creating this knowledge data that are robust, repeatable, and accurate.
- For users of this data to benefit from the querying capabilities of knowledge graphs
- To make the publicly available aspects of this clinical study data accessible to other agents whilst simultaneously handling issues of privacy and confidentiality where necessary.

Research Questions

The research questions this study will aim to answer are:

- Can ontologies provide a useful unified view of the complex use case presented by the Perspectum study data?
- Can the data in this domain, that is, metrics related to imaging clinical studies, be put into context using a knowledge graph, that is, are the hierarchies and relationships at the knowledge graph level useful to profile the metrics and organ data?

- Can we (semi)automate and streamline the creation of the ontology and knowledge graph via templates and other recently researched methods?
- Is there a feasible alternative to the creation of a new Semantic data source, that is, the use of an ontology and mappings to access the data in a non-Semantic form?

Outputs

This project will be a design and build study. Perspectum will provide a collection of synthesised datasets, representing the study metrics and modelled on reasonable, realistic patient scenarios from the previously published research. The data will have largely tabular form.

The major outputs of this project will be:

- An ontology of the data in scope.
- A set of software modules to take the supplied tabular data and create as Semantic data stores.
- A means of accessing and interrogating the created Sematic data to allow extraction and analysis.
- An evaluation of the methods used and artefacts produced.

Scope

The initial scope will focus on a small subset of the available patient data, initially the metrics of a single organ, the liver, and a subset of other attributes (eg. demographics, baseline measures) agreed with the co-supervisor. This will allow the analysis/design/build lifecycle to be completed more quickly, providing evidence and results of the study earlier and more reliably to allow feedback and revision. Further iterations are intended with an expanded scope.

Beneficiaries

Perspectum: this study will demonstrate the feasibility of storing this domain of data with Semantic Technologies. They will have a report of the options explored in building this artefact and a basis to take the product further if the benefits of doing so are sufficient.

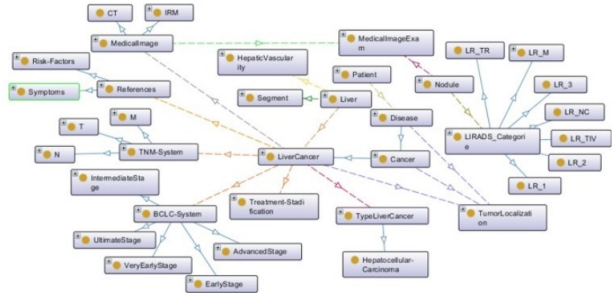
City University Research: this study will be an addition to the literature on domains that can benefit from Semantic Web Technologies and a documentation of the strengths and weaknesses found in the methods and tools used for this domain.

The researcher: this study will allow me to gain experience of an academic research project and in formulating a Semantic Web solution to a real-world problem, exploring the options available.

Critical Context

Ontologies are already widely used in the bio-medical field but also in a wide range of other domains, as diverse as offshore oil platform building (Skjæveland et al, 2018b) and digital humanities (Shimizu et al, 2020a), with documented benefits.

A literature search did not find any specific examples of the Perspectum use case but there are examples of ontologies related to generic human clinical study data (Tu et al, 2009) and to other liver diseases (Figure 1, Messaoudi et al, 2019) which will provide useful input to the project ontology.



In the bio-medical domain, the Open Biological and Biomedical Ontologies (OBO) Foundry exists to co-ordinate the accurate evolution of ontologies to support biomedical data integration (The OBO Foundry, 2021); their ontologies and principles will provide useful guidance.

Figure 1. Visualisation of an Ontology for modelling liver cancer (Messaoudi et al, 2019)

Semantic Web technologies are intended to increase the availability and reuse of data. The FAIR principles (Wilkinson et al, 2019) detail 15 guidelines intended to increase the ability of machines to discover and use data by improving its ability to be Findable, Accessible, Interoperable and Reusable (FAIR).

Recent literature explores ways to put the FAIR principles into practice. Jacobsen (2020) describes implementation considerations, whilst Wilkinson (2017) recommends a combination of W3C's Linked Data Platform, RDF mapping language and triple pattern fragments to create an infrastructure for data and metadata that meets the FAIR principles.

There are recent examples of appropriate methodologies with which to develop Semantic data stores. ROBOT (Jackson, 2019) is a suite of tools to automate the related ontology development tasks of requirements gathering, ontology development and publishing, feedback, and deployment. Sequeda et al (2019) document an iterative methodology, 'pay-as-you-go', as a process of ontology creation and mapping from a relational database, scoped and driven by the prioritised business questions that are to be answered from the data. Although this is focussed on large databases, the described tasks of Knowledge Capture, Knowledge Implementation and Self-Service Analytics, and their supporting documents, may provide a useful framework to plan and scope any modelling problem into smaller sub-tasks.

As an alternative to the creation of new Semantic data stores, many recent papers, including Sequeda et al (2019) and Kharlamov et al (2015), also describe Ontology-Based Database Access (OBDA) as a means of accessing the data in non-Semantic structures by means of an Ontology and mappings to the underlying sources. If time allows, this project will consider this as an alternative strategy.

Re-use of ontologies is an important facet of Semantic Technologies and modular development has been examined as a means of encouraging this. Shimizu et al (2020b) describe the Modular Ontology Modelling (MOMo) methodology that uses modular development and reuse of design patterns. Particularly emphasised is the use of graphical schema diagrams to elicit knowledge from domain experts. Also discussed as important to re-usable ontologies are concepts and relationships meaningful to domain experts, and the following of established modelling principles.

Templates are suggested to simplify ontology development. Skjæveland et al (2018a) introduce OTTR, a template-based means of creating ontologies, abstracting repeated patterns in the knowledge structure to simplify the creation and maintenance. Similarly, ROBOT (no date) provides a method of using templates to convert from tables to OWL format that may also be applicable to conversion of tabular data.

As well as examining the use of Semantic technologies in MRI clinical study data, this project will evaluate the above approaches in increasing re-use and improving accuracy and timeliness of the processes.

Approaches

This project will be undertaken as a Design and Build study. It will begin with a review of the relevant literature, for similar domain ontologies and for the current state of Semantic Web Technologies, with evaluation of relevant tools and methodologies. Following this review, tools and methods will be chosen and used in more detailed task planning. The design and build will be carried out over several iterations, each tackling a small, defined subset of the study data attributes so that delivery of useful artefacts can be completed more quickly, and lessons can be learned and applied to subsequent iterations.

It will be supervised by Dr Ernesto Jiménez-Ruiz, Lecturer in Artificial Intelligence, City University of London and by Dr Valentina Carapella, Data Scientist, Perspectum. Valentina is also the domain expert and source of knowledge for the data.

Regular meetings will be scheduled with the Supervisors to monitor progress and receive feedback.

Methods

- The initial plan will employ aspects of the 'pay-as-you-go' methodology (Sequeda et al, 2019) to structure the project tasks (Figure 2).
- This project will aim to produce artefacts which abide by the FAIR guidelines and will explore suggestions in recent literature (Wilkinson, 2019), (Jacobsen, 2020) to evaluate the success of this objective.
- The process of knowledge graph creation will be built in modules to be reusable and automated as far as possible.

Tools

- Ontology modelling will be carried out in Protégé.
- Newly created code modules will be in Python.
- The literature review will identify which other tools and methodologies may be appropriate and evaluate their usefulness for the project.

Tasks

An overview of the proposed tasks is shown in Figure 2.

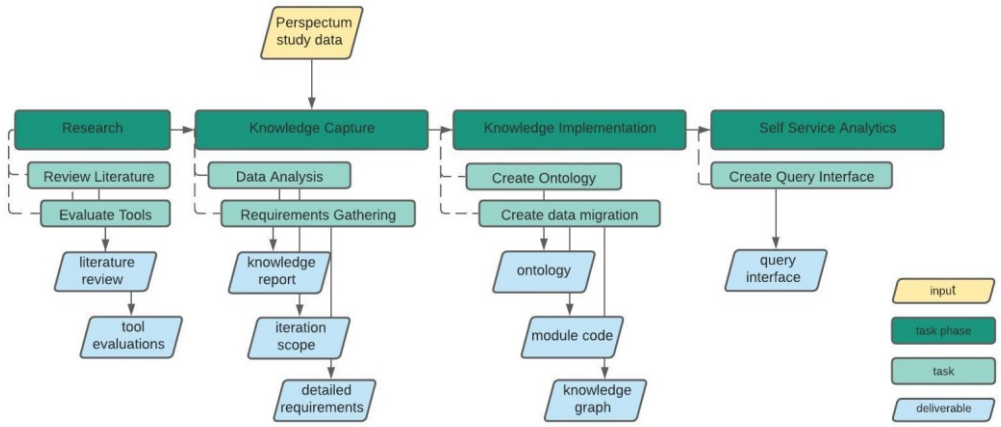


Figure 2. Project Tasks and Deliverables

Search and Review Literature

Explore the use of ontologies and semantic modelling in this or similar domains: MRI, biomarkers, clinical trial data, OBO/OBI. Investigate guidelines and protocols to be used to conform to current best practice.

Evaluate Tools

Investigate costs and benefits of prospective methodologies and tools from the literature to evaluate which would be of benefit in the project. To include, but not limited to, GraphDB (2021), AllegroGraph (2021), OTTR (Skjæveland et al, 2018a), ROBOT (Jackson, 2019), (ROBOT, no date) and MODL (Shimizu, 2019).

Data Analysis & Requirements Gathering

The synthetic study data supplied from Perspectum will be analysed in conjunction with domain experts and documented in a Knowledge Report. Detailed descriptions and features of each data item will be included (meaning, usage, data types, valid values, optionality, etc) to give a thorough understanding and provide input to the semantic metadata of the eventual Ontology.

When the data and requirements are understood and documented, the design and build of the deliverables will be carried out in iterative development cycles (see Work Plan), beginning with the initial scope of concepts and data.

For each Iteration:

Create Ontology

Model the knowledge of the data in scope in an ontology using Protégé (2021), employing validation and reasoning.

Create Data Migration

Design and build a data mapping of the in-scope source study data to the created ontology entities. Code modules to create the new Semantic data store. The outcome of the tool evaluation tasks will inform a more detailed plan of the work in this phase.

Create Query Interface

Using tools identified in the Tool Evaluation task, create an interface to allow domain experts and other users to query the data.

Evaluate Iteration

At the completion of an iteration, a query interface should be delivered to the domain expert for evaluation. Full feedback of results will inform the next iteration.

Evaluation

To assess the success of the project, we propose a series of checklists so that the evaluation is as quantitative as possible, to include the following:

- Elicit a collection of analytical questions that Perspectum would like to answer from the data and use as evaluation criteria.
- Assess the created data on each of the 15 FAIR principles. A checklist of implementation guidelines will be created to which the products should conform.
- Assess the created data on the linked data principles defined by Tim Berners-Lee (2009).

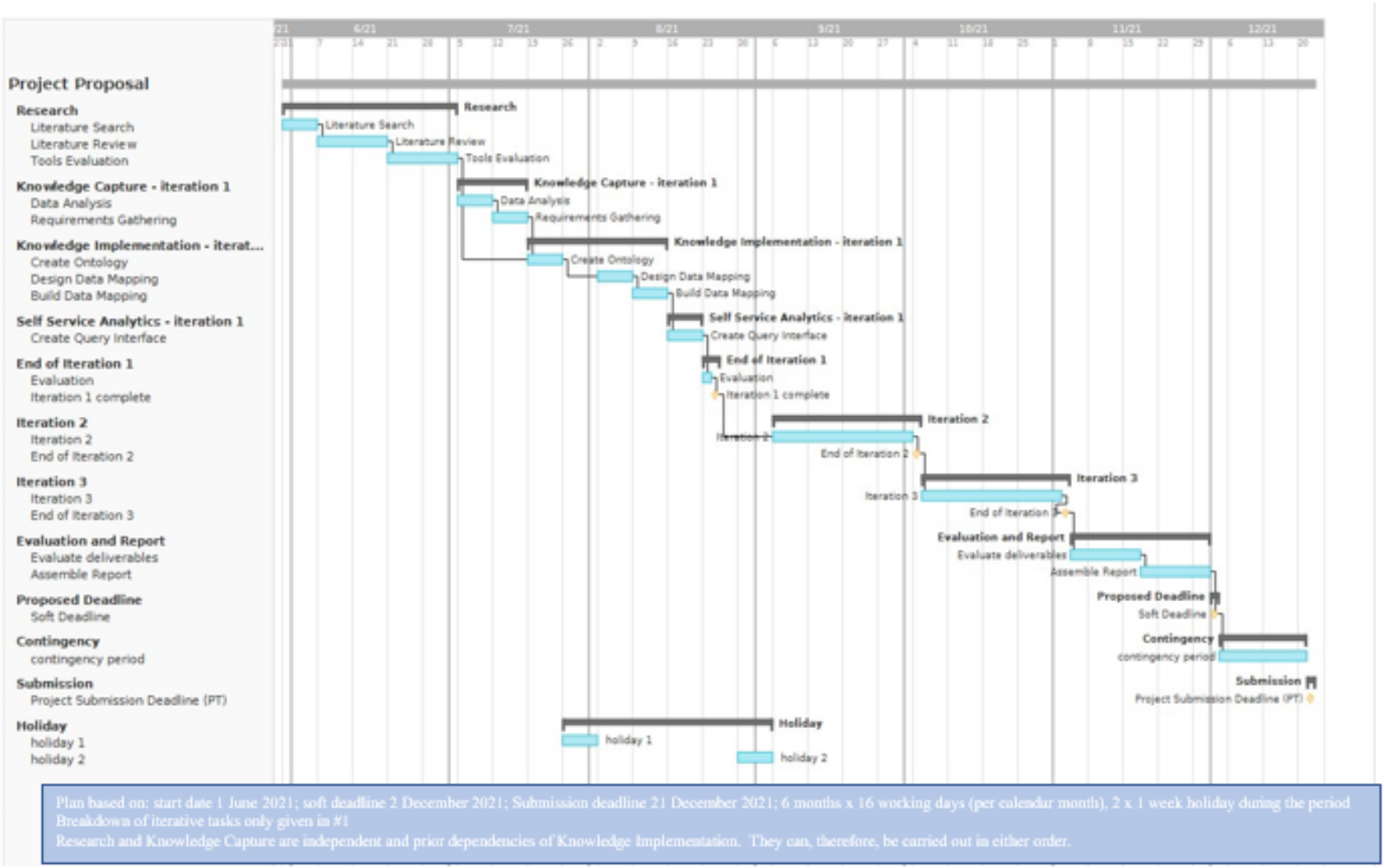
Ethical, Legal and Professional issues

All external research and software will be referenced.

There are no additional participants in this study and all data used is synthesised so there are no risks to the identification of individuals nor use of sensitive corporate information.

Part A of the CS Research Ethics Review Form is included at the end of this proposal and all questions (1-18) are answered ‘NO’ so the project will not require referral to an ethics committee.

Work Plan


Project Proposal
Research
Literature Search
Literature Review
Tools Evaluation
Knowledge Capture - iteration 1
Data Analysis
Requirements Gathering
Knowledge Implementation - iterat...
Create Ontology
Design Data Mapping
Build Data Mapping
Self Service Analytics - iteration 1
Create Query Interface
End of Iteration 1
Evaluation
Iteration 1 complete
Iteration 2
Iteration 2
End of Iteration 2
Iteration 3
Iteration 3
End of Iteration 3
Evaluation and Report
Evaluate deliverables
Assemble Report
Proposed Deadline
Soft Deadline
Contingency
contingency period
Submission
Project Submission Deadline (PT)
Holiday
holiday 1
holiday 2
6/21
7/21
8/21
9/21
10/21
11/21
12/21
Plan based on: start date 1 June 2021; soft deadline 2 December 2021; Submission deadline 21 December 2021; 6 months x 16 working days (per calendar month), 2 x 1 week holiday during the period
Breakdown of iterative tasks only given in #1
Research and Knowledge Capture are independent and prior dependencies of Knowledge Implementation. They can, therefore, be carried out in either order.

Risk Register

1=very high 2=high 3=medium 4=low 5=very low

Risk	Likelihood	Impact	Mitigation
The project tasks take longer than anticipated and runs out of time.	2	1	Work in an 'Agile' way to ensure early and iterative deliverables. Scope well, plan tasks in detail and monitor progress closely during the project. Include contingency in the project plan and aim for a project completion milestone before the University submission date.
I am not familiar enough with technology required and it takes a greater proportion of the project to become proficient.	3	2	Identify what is required in detail in the plan and allocate (pre-)project tasks to learning/familiarisation. Monitor task progress during the project and, if necessary, revisit the scope of new tools and methods.
The supervisors become unavailable.	3	3	Elicit as much information from supervisors at the beginning of the project as part of the requirements gathering process; Ensure the success criteria are academic and related to research stakeholders as well as Perspectum-related; if required information is not available, make and document assumptions to continue with the project.
Computer failure, data loss	4	2	Use online version management software for code, use cloud storage for other work. Create a disaster recover plan.
I have no experience in a bio-medical domain and the specialist knowledge required might make it difficult to understand sufficiently to design the ontology	3	1	Allocate time to perform background reading in this domain and Perspectum's work. Understand and document the specialist vocabulary. Include research from related domains in the literature review.

References

AllegroGraph (2021) AllegroGraph Overview. Available at: https://allegrograph.com/products/allegrograph/ (accessed: 15 April 2021)

Dennis, A. et al. (2020) 'Multi-organ impairment in low-risk individuals with long COVID', medRxiv, p. 2020.10.14.20212555. doi: 10.1101/2020.10.14.20212555.

Hitzler, P., Krötzsch, M. & Rudolph, S.(.s. 2010, Foundations of Semantic Web technologies, CRC Press, Boca Raton.

Jackson, R.C. et al. 2019, "ROBOT: A Tool for Automating Ontology Workflows", BMC bioinformatics, vol. 20, no. 1, pp. 407-407.

Jacobsen, A. et al. 2020, "FAIR principles: interpretations and implementation considerations", Data Intelligence, vol. 2, no. 1-2, pp. 10-29.

Kharlamov, E. et al. 2015, "Ontology Based Access to Exploration Data at Statoil" in Springer International Publishing, Cham, pp. 93-112.

Messaoudi, R. et al. (2019) 'Ontology-Based Approach for Liver Cancer Diagnosis and Treatment', *Journal of Digital Imaging*, 32(1), pp. 116–130. doi: 10.1007/s10278-018-0115-6

Ontotext (2021) GraphDB. Available at: https://www.ontotext.com/products/graphdb/ (accessed: 15 April 2021)

Perspectum (2021) Our Company. Available at: https://perspectum.com/ (accessed: 16 April 2021)

Protégé (2021) About. Available at: protégé (stanford.edu) (accessed 30 April 2021)

ROBOT (no date) Template. Available at: http://robot.obolibrary.org/template (accessed: 30 April 2021)

Sequeda, J.F. et al. 2019, "A Pay-as-you-go Methodology to Design and Build Enterprise Knowledge Graphs from Relational Databases" in Springer International Publishing, Cham, pp. 526-545.

Shimizu, C. et al (2020) Modular Ontology Modeling. Available at: http://www.semantic-web-journal.net/content/modular-ontology-modeling (accessed: 30 April 2021)

Shimizu, C., Hitzler, P. et al. 2020, "The enslaved ontology: Peoples of the historic slave trade", Web semantics, vol. 63, pp. 100567.

Shimizu, C., Hirt, Q. & Hitzler, P. 2019, "MODL: A Modular Ontology Design Library", .

Skjæveland, M. et al. "Semantic Material Master Data Management at Aibel." International Semantic Web Conference (2018).

Skjæveland, M.G., Lupp, D.P., Karlsen, L.H. & Forssell, H. 2018, "Practical Ontology Pattern Instantiation, Discovery, and Maintenance with Reasonable Ontology Templates" in Springer International Publishing, Cham, pp. 477-494.

The OBO Foundry (2021) The Open Biological and Biomedical Ontology (OBO) Foundry. Available at: www.obofoundry.org (accessed: 23 April 2021)

Tim Berners-Lee (2009) Linked Data. Available at: https://www.w3.org/DesignIssues/LinkedData.html (accessed: 30 April 2021)

Tu, S. *et al.* (2009) 'OCRe: An Ontology of Clinical Research'.

Wilkinson, M.D. et al 2019, "The FAIR Guiding Principles for scientific data management and stewardship (vol 15, 160018, 2016)", Scientific data, vol. 6.

Wilkinson, M.D. et al 2017, "Interoperability and FAIRness through a novel combination of Web technologies", PeerJ. Computer science, vol. 3, (April 2017) pp. e110. https://doi.org/10.7717/peerj-cs.110 (accessed: 23 April 2021)

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

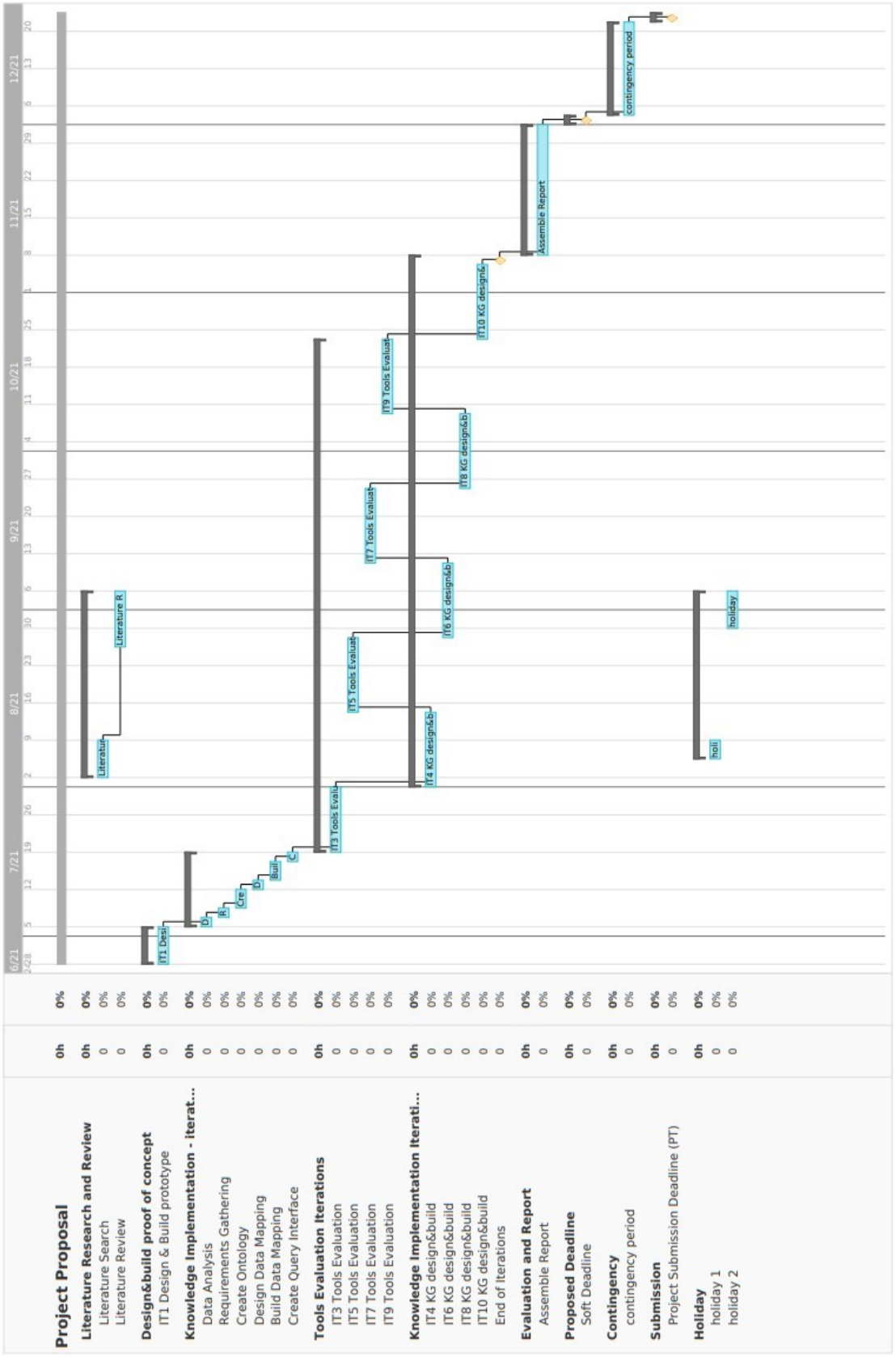
PART B: Ethics Proportionate Review Form. Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk. The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online -		*Delete as appropriate*

https://ethics.city.ac.uk/		
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London?	**NO**

	For example, students studying on a particular course or module. *If yes, then approval is also required from the Head of Department or Programme Director.*	
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**

Appendix B Revised Project Plan



Appendix C Knowledge Report[45]

Business Questions	Answers
What: What are the business questions? What is the business problem?	Perspectum creates and holds datasets relating to multiple clinical studies. They consist of metrics relating to patients, their health and various biomarkers and imaging results gathered for the study. There is currently no unified way to answer complex questions of this data. The data is currently held in multiple spreadsheets and is not widely accessible, nor easy to understand or query. Perspectum would like this data to be more accessible and structured in a way that allows non-experts to formulate and answer queries. For this project, the results of a clinical study of long Covid patients will be used to evaluate the usefulness of Ontologies and Knowledge Graphs for this domain. Initially, the aim is to make this data available within Perspectum but in the longer term this could be extended to stakeholders outside the company. Specific business questions could be, eg., show all data attributes, including the model of scanner used, for female patients, in a specific age range, with BMI 30-40 whose Liver cT1 > 800ms.
Why: Why do we need to answer these questions? What is the motivation?	Perspectum scientists wish to analyse previously gathered patient imaging and study data for various reasons. They may need to analyse this data to, for example, draw new biomedical conclusions from previous studies, to identify sub-cohorts of patient volunteers or examine trends in imaging data over time. At each stage of gathering and recording imaging study data, there are attributes of how each metric is obtained, stored or accepted. The study Metrics need to be accountable, reliable and reproducible and the data analyses should include attributes defining these factors.
Who: Who produces the data? Who will consume the data? Who is involved?	Patient and imaging metrics are gathered by the medical side of Perspectum. They conduct patient questionaires and assessments, perform MRI scans and record data on imaging success and quality, processing from patient assessment to storage of analysis output. Data Scientists perform additional computations on the output of this analysis to create validated metrics for each participant, producing final datasets on which statistics and analyses can be run. Biomedical Scientists can then use the final datasets to answer their questions.
How: How is this the business question answered today, if at all?	Currently Perspectum Data Scientists are the only people with the knowledge to manipulate and analyse the output of the medical assessments and must regularly spend time combining datasets, creating queries and producing results to answer questions from other areas of the company. This involves valuable time and resource and repetition of effort.
Where: Where are the data sources required to answer the business questions?	Data Scientists create and own the input spreadsheets that will be used to populate the Knowledge Graph(s).
When: When will the data be consumed? Real-time? Daily? Update criteria?	Study, patient and imaging data will be available to update the KG on an ad-hoc basis. It is assumed that data will be provided in pre-defined spreadsheet formats and that load modules should be able to create and add to the various sub graphs that will be defined. These requirements will be analysed in further detail as the project progresses. Consumption and analysis of the final data will be required on request.

[45] https://docs.google.com/spreadsheets/d/16A43SyfZajoQdtNoMj7_KMihXr3w3bdm/edit#gid=1975891667

Name	Definition	ID (URI)	Source column	Source File/sheet
Patient	A human participant in the Perspectum long COVID study. For the purposes of this ontology, the patient is anonymously identified by a patient identifier, Pnnn, where n is a unique number.	URI_Patient_%patient id%, eg. URI_Patient_P275	Patient_ID	'TabularData'
ScanVisit	A visit made by a patient for the purpose of taking part in an MRI scan to gather metrics for the study data. A scan visit is defined as a single visit for a single patient and allows identification of multiple values of a single metric type for a single patient. A visit is currently identified by a sequential number currently given in the test data.	URI_Visit_%patient_id%_%visitID%, eg.URI_Visit_P005_2	Scan visit	'TabularData'
QuantitativeMetric	A quantitative metric gathered for the study. The metric may either be scan or patient related. Example metrics are Patient Age or Liver cT1.	see sub-types, LivercT1, etc	liver_cT1, liver_PDFF, liver_T2star	'TabularData'
LivercT1	Corrected T1 is an MRI derived liver biomarker providing a metric relating to inflammation and fibrosis. T1-relaxation time (measured in milliseconds) is a fundamental parameter in MRI relating to the interaction and energy exchange between the excited hydrogen atoms (usually in water) and the surrounding tissue structure. A sub class of Quantitative Metric. T1 measurements differ depending on magnetic field strength and MR manufacturer.	URI_LivercT1_%metricID%, eg. URI_liver_cT1_1000	liver_cT1	'TabularData'
LiverPDFF	Liver Protein Density Fat Fraction (PDFF) is a measure of hepatic fat. It is an MRI derived biomarker, given as a percentage (%) , which uses the difference in resonance frequencies of the protons in water and fat to provide estimates of tissue fat fraction, given by fat / (water + fat). PDFF has been widely shown to corelate with the degree of hepatic steatosis with a cut-off of 5% being indicative of Non-Alcoholic Fatty Liver Deasease (NALFD). A sub class of Quantitative Metric. (https://perspectum.com/media/1361/understanding-pdff.pdf)	URI_liver_PDFF_%metricID%, eg. URI_liver_PDFF_999	liver_PDFF	'TabularData'
LiverT2Star	T2* is an MRI derived biomarker that can be used in the assessment of hepatic iron overload, or hemosiderosis. Measuring iron load is important in assessing liver health and the relationship between T2* and liver iron concentration (mg Fe/g dry weight tissue) has been well validated. A sub class of Quantitative Metric. T2* is a time constant (measured in milliseconds) describing the decay of an MRI signal and can indicate iron deposits in the liver. T2* is field strength dependant.	URI_LiverT2Star_%metricID%, eg. URI_LiverT2Star_1002	liver_T2star	'TabularData'
PatientBMI	The Body Mass Index of a patient. BMI is calculated as weight in KG squared / height in metres. A sub class of Quantitative Metric.	URI_PatientBMI_%patient id%_%visit id%, eg.URI_PatientBMI_P005_2	BMI	TabularData'

PatientAge	The age of a patient, in years, recorded at a Scan Visit. A sub class of Quantitative Metric.	URI_PatientAge_%patient id%_%visit id%, eg.URI_PatientAge_P005_2	Age	TabularData'
PatientMetric	A quantitative metric gathered for the study and related to a patient	see sub-types, LivercT1, etc	liver_cT1, liver_PDFF, liver_T2star	'TabularData'
MRIScannerModel	An MRI (Magnetic resonance imaging) scanner is a large tube that contains powerful magnets and can be used to examine almost any part of the body. Magnetic resonance imaging (MRI) is a type of scan that uses strong magnetic fields and radio waves to produce detailed images of the inside of the body. The results of an MRI scan can be used to help diagnose conditions, plan treatments and assess how effective previous treatment has been. (https://www.nhs.uk/conditions/mri-scan/)	URI_Scanner_%Manufacturer%_%Model%, eg. URI_Scanner_Philips_ModelA	Scanner	'ScannerTypes'
ScannerManufacturer	A company that manufactures and sells MR scanners for human diagnostic use.	URI_ScannerManufacturer_%Name%, eg. URI_ScannerManufacturer_Philips	Scanner Manufacturer	'ScannerTypes'
ScanMetric	A quantitative metric gathered for the study and derived from an MRI scan.	see sub-types, LivercT1, etc	liver_cT1, liver_PDFF, liver_T2star	'TabularData'
ScannerFieldStrength	The measure of the strength of the magnet used in an MRI Scanner.	URI_ScannerFieldStrength_%MRIscannerModelName% eg. URI_ScannerFieldStrength_Siemens_ModelA	Scanner Field Strength	'ScannerTypes'

PatientSex	The biological sex of a patient, valid values 'Male'/'Female''.	PatientSex	Patient	Sex	'TabularData'	string	yes - assumed	assumed no default value
FieldStrengthValue	The field strength of the magnet used in an MRI Scanner, measured in teslas (T) Hospitals routinely use machines with field strengths of 1.5 T or 3 T, but ultra-high-field scanners are on the rise.	FieldStrengthValue	ScannerFieldStrength	Scanner Field Strength	'ScannerTypes'	double	no	
FieldStrengthUnit	The unit of measure of the scanner field strength. Usually tesla (T)	FieldStrengthUnit	ScannerFieldStrength	Unit	'ScannerTypes'	string	no	
MetricValue	The numeric value of a Quantitative Metric.	MetricValue	Quantitative Metric	liver_cT1, liver_PDFF, liver_T2star	'TabularData'	double	no	

Name	Definition	id (URI)	From Concept	Source Column (from)	Source File/sheet	To Concept	Source Column (to)
isAttendedBy	The attendance at a scan visit of a patient in the study. A scan visit may be attended by many patients.	psp:isAttendedBy	ScanVisit	Scan visit	'TabularData'	Patient	Patient_ID
attendsVisit	The attendance of a patient at a scan visit. A patient may attend many visits.	psp:attendsVisit	Patient	Patient_ID	'TabularData'	ScanVisit	Scan visit
hasPatientMetric	to show a Patient has a particular Quantitative Metric recorded	psp:hasPatientMetric	Patient	Patient_ID	'TabularData'	Quantitative Metric	liver_cT1, liver_PDFF, liver_T2star
isMetricForPatient	to show a Quantitative Metric is recorded for a particular Patient.	psp:isMetricForPatient	Quantitative Metric	liver_cT1, liver_PDFF, liver_T2star	'TabularData'	Patient	Patient_ID
isMetricForVisit	to show a Quantitative Metric is recorded at a particular Scan Visit.	psp:isMetricForVisit	Quantitative Metric	liver_cT1, liver_PDFF, liver_T2star	'TabularData'	ScanVisit	Scan visit
hasVisitMetric	to show a scan visit has resulted in a particular Quantitative Metric	psp:hasVisitMetric	ScanVisit	Scan visit	'TabularData'	Quantitative Metric	liver_cT1, liver_PDFF, liver_T2star
usedInVisit	to show an MRI Scanner Model is used in a particular Scan Visit.	psp:usedInVisit	MRIScannerModel	Scanner	'TabularData'	ScanVisit	Scan visit
usesScannerModel	to show a Scan Visit used a particular MRI Scanner Model	psp:usesScannerModel	ScanVisit	Scan visit	'TabularData'	MRIScannerModel	Scanner
isMadeBy	to show that a particular MRI Scanner Model is made and sold by a Manufacturer.	scn:isMadeBy	MRIScannerModel	Scanner Model	'ScannerTypes'	ScannerManufacturer	Scanner Manufacturer
isMakerOf	to show that a Manufacturer makes and sells a particular MRI Scanner Model.	scn:isMakerOf	ScannerManufacturer	Scanner Model	'ScannerTypes'	MRIScannerModel	Scanner Model
hasFieldStrength	to show that an MRI Scanner Model has a magnetic Field Strength	scn:hasFieldStrength	MRIScannerModel	Scanner Model	'ScannerTypes'	ScannerFieldStrength	Scanner Field Strength

isFieldStrengthForScanner	to show that a Field Strength is pertinent to an MRI Scanner Model	scn:isFieldStrengthForScanner	ScannerFieldStrength	Scanner Field Strength	'ScannerTypes'	MRIScannerModel	Scanner Model

Appendix D Evaluation – FAIR data

An assessment of the extent to which the project ontology and data meet the FAIR principles for Findability, Accessibility, Interoperability and Reusability.

Referencing against dint_r_00024.pdf (soton.ac.uk) implementation principles

FAIR ID	FAIR Principle	Met?	Required
F1	(meta)data are assigned a globally unique and persistent identifier	Partly	Identifiers not necessarily persistent. DOI required?
F2	data are described with rich metadata	Yes	Check this to assess: S.-A. Sansone. FAIRsharing as a community approach to standards, repositories and policies.
F3	metadata clearly and explicitly include the identifier of the data it describes	~~No~~ Yes	Look at isDefinedBy property. Added to ontologies and KGs
F4	(meta)data are registered or indexed in a searchable resource	No	'Current challenges are numerous, significantly limiting, and largely outside of the control of the average data provider. ' Eg Google Dataset Search exists but has limitations.
A1	(meta)data are retrievable by their identifier using a standardized communications protocol	Yes	
A1.1	the protocol is open, free and universally implementable	Yes	HTTP
A1.2	the protocol allows for an authentication and authorization procedure, where necessary	NA	Not currently applicable
A2	metadata are accessible, even when the data are no longer available	NA	Not currently applicable
I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	Yes	RDF used
I2	(meta)data use vocabularies that follow FAIR principles	Yes	OWL, RDFS
I3	(meta)data include qualified references to other (meta)data	Yes	Currently linked to QUDT UoM ontology and Vocabulary. To look at other relevant external data that would be useful to link to.
R1	(meta)data are richly described with a plurality of accurate and relevant attributes	Yes	'the focus of R1 is to enable machines and humans to assess if the discovered resource is appropriate for reuse,'
R1.1	(meta)data are released with a clear and accessible data usage license	NA	Add licence meta-data 'There are good reasons for choosing a CC0 license for data'
R1.2	(meta)data are associated with detailed provenance	No – to add	'Several early tools are under development to make the construction of FAIR metadata

FAIR ID	FAIR Principle	Met?	Required
			easier, including for instance CEDAR□20, CASTOR□21'
R1.3	(meta)data meet domain-relevant community standards	No	'Several disciplinary communities have defined Minimal Information Standards describing most often the minimal set of metadata items required to assess the quality of the data acquisition and processing and to facilitate reproducibility. Such standards are a good start, noting that true (interdisciplinary) reusability will generally require richer metadata. For a list of such standards, consult FAIRsharing□23 'An example of minimal information standards is the MIAME standard [27], and various metadata profiles have been defined on top of specifications (e.g. various DCAT profiles)'

Appendix E Python Code

All Python code submitted separately in:

JudithGrievesAllCode.txt

And individual Python modules available in:

https://github.com/JudithGrieves/City-MSc-Project/tree/main/code

Appendix F RML Mapping File

All RML code submitted separately in:

JudithGrievesAllCode.txt

And individual RML files available as *.rml.ttl in:

https://github.com/JudithGrieves/City-MSc-Project/tree/main/data

Appendix G Ontology

All ontology extract code submitted separately in:

JudithGrievesAllCode.txt

And Ontology extract files available in:

https://github.com/JudithGrieves/City-MSc-Project/tree/main/ontology

Appendix H Test input data & system test results

Input test data

patient_id	scan_visit	scanner_model	liver_cT1	liver_T2st	liver_PDFI	Test_liver	sex	BMI	age
P106	2	Siemens_ModelA	717.0993	32.18719	5.385632	1714.157	F	21	36
P106	1	Siemens_ModelA	865.1961	31.52439	3.569331	1118.163	F	21	36
P107	2	Siemens_ModelA	723.4436	17.93461	2.394296	1652.979	F	26	48
P107	1	Siemens_ModelA	776.7396	45.46637	not quant	1254.698	F	26	48
P108	2	Siemens_ModelA	871.4954	21.38349	1.778921	1299.88	M	23	48
P108	1	Siemens_ModelA	812.6968	23.99314	4.728744	1487.152	M	23	48
P109	2	Siemens_ModelA	686.1604	26.25378	11.80486	1424.074	F	18	26
P109	1	Siemens_ModelA	718.3203	36.48908	not quant	1790.216	F	18	26
P110	2	Siemens_ModelA	597.1664	45.10725	0.634905	2411.419	F	28	59
P110	1	Siemens_ModelA	706.7095	27.96504	4.262299	1364.946	F	28	59
P111	2	Siemens_ModelA	738.7703	26.77315	not quant	682.6534	F	32	35
P111	1	Siemens_ModelA	616.3866	26.64439	4.194251	1346.055	F	32	35
P112	2	Siemens_ModelA	637.2135	13.32374	not quant	1425.883	F	18	45
P112	1	Siemens_ModelA	708.3209	17.37978	2.020771	1317.326	F	18	45
P113	2	Siemens_ModelA	571.0075	43.873	1.65066	1263.546	M	38	45
P113	1	Siemens_ModelA	842.3131	27.60118	not quant	1636.081	M	38	45
P114	2	Siemens_ModelA	694.2239	36.00094	3.723479	1150.475	M	22	19
P114	1	Siemens_ModelA	672.1527	38.23867	6.043647	1539.549	M	22	19
P115	2	Siemens_ModelA	835.0391	32.61723	not quant	1417.152	M	19	50
P115	1	Siemens_ModelA	692.3496	21.07064	0.535121	1588.449	M	19	50
P116	2	Siemens_ModelA	768.6712	27.23201	12.42402	1699.444	NA	26	38
P116	1	Siemens_ModelA	678.9162	21.84803	10.14717	1289.364	NA	26	38
P117	2	GE_ModelA	596.6478	31.5022	7.506843	1393.947	F	32	49
P117	1	GE_ModelA	781.9587	25.67965	not quant	1273.471	F	32	49
P118	2	GE_ModelA	731.814	28.56524	7.352471	1791.114	F	19	48
P118	1	GE_ModelA	793.5688	17.14836	0.341907	1612.001	F	19	48
P119	2	Siemens_ModelB	689.8367	25.48025	9.080187	1347.62	F	24	51
P119	1	Siemens_ModelB	641.1502	23.2118	9.120412	1852.77	F	24	51
P120	2	Siemens_ModelB	714.6421	27.54524	12.7063	1527.504	F	21	33
P120	1	Siemens_ModelB	715.0911	39.41197	not quant	1299.896	F	21	33
P138	2	Siemens_ModelB	553.126	26.13698	not quant	2362.574	M	25	42
P138	1	Siemens_ModelB	722.1959	22.01882	7.983237	1932.622	M	25	42
P139	2	Siemens_ModelB	768.4641	30.21796	3.616453	1774.69	M	27	43
P139	1	Siemens_ModelB	759.4755	28.90918	3.969171	957.7075	M	27	43
P140	2	Siemens_ModelB	839.9307	23.59692	not quant	2074.242	M	27	46
P140	1	Siemens_ModelB	620.6052	31.00944	not quant	2213.455	M	27	46
P141	2	Siemens_ModelB	822.5785	43.83864	5.374415	918.9986	M	22	41
P141	1	Siemens_ModelB	659.658	26.37957	12.45399	1954.999	M	22	41
P152	2	Siemens_ModelB	815.3653	29.06105	5.505948	815.5348	M	32	39
P152	1	Siemens_ModelB	706.189	36.24411	3.856706	1262.062	M	32	39
P153	2	Siemens_ModelB	685.753	27.92067	3.482587	1679.542	M	33	43
P153	1	Siemens_ModelB	655.1345	26.09479	6.89467	1903.258	M	33	43
P154	2	Siemens_ModelB	680.5191	36.13999	10.37628	1591.775	M	30	44
P154	1	Siemens_ModelB	751.3724	30.2511	not quant	1846.071	M	30	44
P155	2	Siemens_ModelB	661.6193	21.73516	not quant	1475.26	F	27	43
P155	1	Siemens_ModelB	716.6004	32.67905	3.957455	1805.628	F	27	43
P156	2	Siemens_ModelB	680.9765	31.14609	4.038765	1580.886	F	25	50
P156	1	Siemens_ModelB	733.6876	38.30529	6.681918	1838.168	F	25	50
P157	2	Siemens_ModelB	736.4036	28.53407	1.41693	1659.472	F	33	52
P157	1	Siemens_ModelB	726.8199	36.28123	6.508205	1816.015	F	33	52
P158	2	Siemens_ModelB	744.8608	31.32931	7.311009	1869.894	F	25	44
P158	1	Siemens_ModelB	833.3216	34.27713	4.30133	1799.667	F	25	44
P206	2	Siemens_ModelB	701.5011	40.78833	0.920196	1260.727	F	28	60
P206	1	Siemens_ModelB	839.155	24.11419	1.19701	1933.087	F	28	60
P207	2	Siemens_ModelB	730.5404	23.09279	9.866365	1323.986	F	20	37
P207	1	Siemens_ModelB	559.5094	32.472	6.736574	1772.967	F	20	37
P208	2	Siemens_ModelB	704.2965	17.12277	10.08663	857.7621	F	29	45
P208	1	Siemens_ModelB	489.5996	14.12555	3.106489	1780.883	F	29	45
P209	2	GE_ModelA	695.0878	37.23095	not quant	1537.928	F	31	57
P209	1	GE_ModelA	542.4716	21.64066	not quant	2187.945	F	31	57
P210	2	GE_ModelA	628.8713	22.55355	9.043956	1589.04	M	20	43
P210	1	GE_ModelA	744.4884	38.57379	13.43601	1662.368	M	20	43
P211	2	GE_ModelA	667.3848	27.20857	5.596022	1397.463	M	20	35
P211	1	GE_ModelA	803.8618	25.54522	not quant	2180.679	M	20	35
P212	2	GE_ModelA	858.8568	35.80312	9.818647	2130.883	M	23	60
P212	1	GE_ModelA	698.2964	14.71976	6.771225	1484.588	M	23	60
P213	2	GE_ModelA	615.8037	20.28282	not quant	1579.524	M	30	39
P213	1	GE_ModelA	792.2131	26.32717	9.485298	1503.374	M	30	39
P214	2	GE_ModelA	697.6609	35.6094	not quant	2043.042	F	16	51
P214	1	GE_ModelA	694.9286	32.70519	5.827178	1498.301	F	16	51
P215	2	GE_ModelA	595.3789	40.08415	4.573978	1398.616	F	22	33
P215	1	GE_ModelA	756.2848	34.48953	not quant	1483.804	F	22	33
P216	2	GE_ModelA	731.9459	23.25285	3.71957	1615.27	F	19	37
P216	1	GE_ModelA	677.6997	19.84416	2.717156	1443.204	F	19	37
P217	2	GE_ModelA	946.7474	28.62632	7.983065	983.8969	F	23	41
P217	1	GE_ModelA	666.6643	27.22353	not quant	1532.875	F	23	41
P218	2	GE_ModelA	640.4763	25.47133	2.725241	723.8591	M	29	40
P218	1	GE_ModelA	791.9395	34.23479	4.771042	1523.097	M	29	40
P219	2	GE_ModelA	662.9019	19.81087	not quant	1708.813	M	21	61
P219	1	GE_ModelA	687.4108	31.25038	10.49379	1588.605	M	21	61
P384	2	GE_ModelA	724.4882	32.3415	6.539466	910.2643	M	27	37
P384	1	GE_ModelA	768.8474	37.62997	6.677259	1651.464	M	27	37
P385	2	GE_ModelA	841.2007	25.95964	4.783474	1895.398	M	18	55
P385	1	GE_ModelA	704.6585	16.91431	11.65083	1117.014	M	18	55
P386	1	Philips_ModelA	695.295	21.04967	9.064205	1202.12	M	20	42
P387	1	Philips_ModelA	667.6842	37.12176	8.740496	1186.964	F	18	47
P388	1	Philips_ModelA	775.2139	30.69624	5.002815	1510.817	F	29	42
P389	1	Philips_ModelA	720.3896	42.48708	5.004033	1540.702	F	38	52
P390	1	Philips_ModelA	710.3076	15.94041	9.101896	2005.443	F	21	62
P391	1	Philips_ModelA	594.4356	29.13008	5.283802	1565.07	F	25	25
P392	1	Philips_ModelA	720.6276	21.00407	11.99599	986.5608	NA	26	55
P393	1	Philips_ModelA	658.8378	29.05351	9.469565	1671.227	M	24	61
P394	1	Philips_ModelA	706.8309	34.98759	13.34469	1451.985	M	23	45
P395	1	Philips_ModelA	820.2613	24.94298	3.966198	1452.196	M	16	35
P400	1	Philips_ModelA	674.6951	25.32081	7.951657	1621.84	M	30	41
P401	1	GE_ModelA	798.16	36.91744	3.922555	960.653	M	31	39
P402	1	GE_ModelA	837.6225	32.65987	7.061598	809.0629	M	26	33
P403	1	GE_ModelA	795.9273	35.40268	not quant	2187.257	M	24	42
P404	1	GE_ModelA	739.4539	31.65766	7.293976	1506.43	M	28	54
P412	1	GE_ModelA	812.4731	30.03382	12.53332	1995.218	F	30	63

Replicated data											
PatientID	Visit	Age	Sex	BMI	livercT1	liverPDFF	liverT2Star	Scanner	Field Strength	Unit	Manufacturer
P106	P106/1	36	Female	21	865.1961088	3.56933051	31.52438818	Siemens_ModelA	1.5	Tesla	Siemens
P106	P106/2	36	Female	21	717.0993199	5.385631551	32.18719455	Siemens_ModelA	1.5	Tesla	Siemens
P107	P107/1	48	Female	26	776.7395897	not quantifiable	45.46637492	Siemens_ModelA	1.5	Tesla	Siemens
P107	P107/2	48	Female	26	723.4435961	2.394295683	17.93460507	Siemens_ModelA	1.5	Tesla	Siemens
P108	P108/1	48	Male	23	812.696841	4.728744032	23.99313931	Siemens_ModelA	1.5	Tesla	Siemens
P108	P108/2	48	Male	23	871.4954244	1.778921052	21.38348531	Siemens_ModelA	1.5	Tesla	Siemens
P109	P109/1	26	Female	18	718.3202775	not quantifiable	36.48907968	Siemens_ModelA	1.5	Tesla	Siemens
P109	P109/2	26	Female	18	686.1604384	11.80486209	26.25377831	Siemens_ModelA	1.5	Tesla	Siemens
P110	P110/1	59	Female	28	706.7094811	4.262298652	27.96504055	Siemens_ModelA	1.5	Tesla	Siemens
P110	P110/2	59	Female	28	597.1664494	0.634905121	45.10725443	Siemens_ModelA	1.5	Tesla	Siemens
P111	P111/1	35	Female	32	616.3865974	4.194250647	26.64438535	Siemens_ModelA	1.5	Tesla	Siemens
P111	P111/2	35	Female	32	738.7703402	not quantifiable	26.77315231	Siemens_ModelA	1.5	Tesla	Siemens
P112	P112/1	45	Female	18	708.3208524	2.020770602	17.379776	Siemens_ModelA	1.5	Tesla	Siemens
P112	P112/2	45	Female	18	637.2135107	not quantifiable	13.32374299	Siemens_ModelA	1.5	Tesla	Siemens
P113	P113/1	45	Male	38	842.3131046	not quantifiable	27.60118132	Siemens_ModelA	1.5	Tesla	Siemens
P113	P113/2	45	Male	38	571.0075321	1.650659613	43.87299637	Siemens_ModelA	1.5	Tesla	Siemens
P114	P114/1	19	Male	22	672.1526938	6.043647126	38.23867182	Siemens_ModelA	1.5	Tesla	Siemens
P114	P114/2	19	Male	22	694.2238754	3.723479303	36.00094185	Siemens_ModelA	1.5	Tesla	Siemens
P115	P115/1	50	Male	19	692.34958	0.53512121	21.07063935	Siemens_ModelA	1.5	Tesla	Siemens
P115	P115/2	50	Male	19	835.0390832	not quantifiable	32.61722748	Siemens_ModelA	1.5	Tesla	Siemens
P116	P116/1	38	Not Appl	26	678.9162427	10.14716849	21.84803167	Siemens_ModelA	1.5	Tesla	Siemens
P116	P116/2	38	Not Appl	26	768.6711705	12.42402256	27.2320114	Siemens_ModelA	1.5	Tesla	Siemens
P117	P117/1	49	Female	32	781.9587384	not quantifiable	25.67965095	GE_ModelA	1.5	Tesla	GE Healthcare
P117	P117/2	49	Female	32	596.6478447	7.506842992	31.50220185	GE_ModelA	1.5	Tesla	GE Healthcare
P118	P118/1	48	Female	19	793.5688241	0.341907008	17.14835887	GE_ModelA	1.5	Tesla	GE Healthcare
P118	P118/2	48	Female	19	731.814041	7.352471472	28.56524313	GE_ModelA	1.5	Tesla	GE Healthcare
P119	P119/1	51	Female	24	641.1502489	9.120411938	23.21179668	Siemens_ModelB	3	Tesla	Siemens
P119	P119/2	51	Female	24	689.836653	9.08018691	25.48024622	Siemens_ModelB	3	Tesla	Siemens
P120	P120/1	33	Female	21	715.0911312	not quantifiable	39.41197313	Siemens_ModelB	3	Tesla	Siemens
P120	P120/2	33	Female	21	714.6420529	12.70629861	27.54524022	Siemens_ModelB	3	Tesla	Siemens
P138	P138/1	42	Male	25	722.1958953	7.983236903	22.01881774	Siemens_ModelB	3	Tesla	Siemens
P138	P138/2	42	Male	25	553.1260353	not quantifiable	26.13697866	Siemens_ModelB	3	Tesla	Siemens
P139	P139/1	43	Male	27	759.4755001	3.969171041	28.90917841	Siemens_ModelB	3	Tesla	Siemens
P139	P139/2	43	Male	27	768.4640647	3.616452992	30.2179583	Siemens_ModelB	3	Tesla	Siemens
P140	P140/1	46	Male	27	620.6052029	not quantifiable	31.00944476	Siemens_ModelB	3	Tesla	Siemens
P140	P140/2	46	Male	27	839.9306614	not quantifiable	23.59691649	Siemens_ModelB	3	Tesla	Siemens
P141	P141/1	41	Male	22	659.6579574	12.45399025	26.37956757	Siemens_ModelB	3	Tesla	Siemens
P141	P141/2	41	Male	22	822.5785479	5.374415343	43.83864489	Siemens_ModelB	3	Tesla	Siemens
P152	P152/1	39	Male	32	706.189026	3.856705761	36.24411357	Siemens_ModelB	3	Tesla	Siemens
P152	P152/2	39	Male	32	815.3653214	5.50594835	29.06105001	Siemens_ModelB	3	Tesla	Siemens
P153	P153/1	43	Male	33	655.1345249	6.894669711	26.09479262	Siemens_ModelB	3	Tesla	Siemens
P153	P153/2	43	Male	33	685.7529812	3.48258722	27.92066952	Siemens_ModelB	3	Tesla	Siemens
P154	P154/1	44	Male	30	751.3724341	not quantifiable	30.25109502	Siemens_ModelB	3	Tesla	Siemens
P154	P154/2	44	Male	30	680.519124	10.37628304	36.13998848	Siemens_ModelB	3	Tesla	Siemens
P155	P155/1	43	Female	27	716.6004326	3.957454659	32.67904545	Siemens_ModelB	3	Tesla	Siemens
P155	P155/2	43	Female	27	661.6192825	not quantifiable	21.73515525	Siemens_ModelB	3	Tesla	Siemens
P156	P156/1	50	Female	25	733.6875822	6.681918009	38.30529239	Siemens_ModelB	3	Tesla	Siemens
P156	P156/2	50	Female	25	680.9764729	4.038764598	31.14608979	Siemens_ModelB	3	Tesla	Siemens
P157	P157/1	52	Female	33	726.8198899	6.508204523	36.28122596	Siemens_ModelB	3	Tesla	Siemens
P157	P157/2	52	Female	33	736.4036314	1.416930326	28.53406597	Siemens_ModelB	3	Tesla	Siemens
P158	P158/1	44	Female	25	833.3215699	4.301329931	34.27712933	Siemens_ModelB	3	Tesla	Siemens
P158	P158/2	44	Female	25	744.8607871	7.311009416	31.3293094	Siemens_ModelB	3	Tesla	Siemens
P206	P206/1	60	Female	28	839.1550289	1.197009769	24.11418677	Siemens_ModelB	3	Tesla	Siemens
P206	P206/2	60	Female	28	701.5010963	0.920195711	40.78832937	Siemens_ModelB	3	Tesla	Siemens
P207	P207/1	37	Female	20	559.5094463	6.7365737	32.47200383	Siemens_ModelB	3	Tesla	Siemens
P207	P207/2	37	Female	20	730.5404498	9.866364603	23.09279326	Siemens_ModelB	3	Tesla	Siemens
P208	P208/1	45	Female	29	489.5996381	3.106489446	14.1255505	Siemens_ModelB	3	Tesla	Siemens
P208	P208/2	45	Female	29	704.2965189	10.08663148	17.12277307	Siemens_ModelB	3	Tesla	Siemens
P209	P209/1	57	Female	31	542.4715682	not quantifiable	21.64066027	GE_ModelA	1.5	Tesla	GE Healthcare
P209	P209/2	57	Female	31	695.0877566	not quantifiable	37.23095226	GE_ModelA	1.5	Tesla	GE Healthcare
P210	P210/1	43	Male	20	744.488429	13.43601452	38.57379192	GE_ModelA	1.5	Tesla	GE Healthcare
P210	P210/2	43	Male	20	628.8713234	9.043955953	22.55354557	GE_ModelA	1.5	Tesla	GE Healthcare
P211	P211/1	35	Male	20	803.8617801	not quantifiable	25.54522039	GE_ModelA	1.5	Tesla	GE Healthcare
P211	P211/2	35	Male	20	667.3848256	5.596022468	27.20857213	GE_ModelA	1.5	Tesla	GE Healthcare
P212	P212/1	60	Male	23	698.2964268	6.77122526	14.7197642	GE_ModelA	1.5	Tesla	GE Healthcare
P212	P212/2	60	Male	23	858.8567633	9.818647081	35.80312352	GE_ModelA	1.5	Tesla	GE Healthcare
P213	P213/1	39	Male	30	792.2131306	9.485298348	26.32717244	GE_ModelA	1.5	Tesla	GE Healthcare
P213	P213/2	39	Male	30	615.8037066	not quantifiable	20.2828157	GE_ModelA	1.5	Tesla	GE Healthcare
P214	P214/1	51	Female	16	694.9285975	5.827178417	32.7051945	GE_ModelA	1.5	Tesla	GE Healthcare
P214	P214/2	51	Female	16	697.6608893	not quantifiable	35.6094022	GE_ModelA	1.5	Tesla	GE Healthcare
P215	P215/1	33	Female	22	756.2848011	not quantifiable	34.48952825	GE_ModelA	1.5	Tesla	GE Healthcare
P215	P215/2	33	Female	22	595.3788578	4.573978031	40.08414527	GE_ModelA	1.5	Tesla	GE Healthcare
P216	P216/1	37	Female	19	677.6997038	2.717155608	19.84416188	GE_ModelA	1.5	Tesla	GE Healthcare
P216	P216/2	37	Female	19	731.9458984	3.719570411	23.25284656	GE_ModelA	1.5	Tesla	GE Healthcare
P217	P217/1	41	Female	23	666.6642687	not quantifiable	27.22352804	GE_ModelA	1.5	Tesla	GE Healthcare
P217	P217/2	41	Female	23	946.7473675	7.983064937	28.62631699	GE_ModelA	1.5	Tesla	GE Healthcare
P218	P218/1	40	Male	29	791.9395219	4.771042386	34.23479084	GE_ModelA	1.5	Tesla	GE Healthcare
P218	P218/2	40	Male	29	640.476299	2.725241318	25.47132545	GE_ModelA	1.5	Tesla	GE Healthcare
P219	P219/1	61	Male	21	687.4108323	10.49378927	31.2503842	GE_ModelA	1.5	Tesla	GE Healthcare
P219	P219/2	61	Male	21	662.9019182	not quantifiable	19.81086504	GE_ModelA	1.5	Tesla	GE Healthcare
P384	P384/1	37	Male	27	768.8474283	6.677259148	37.62997035	GE_ModelA	1.5	Tesla	GE Healthcare
P384	P384/2	37	Male	27	724.4882198	6.539466095	32.34150005	GE_ModelA	1.5	Tesla	GE Healthcare
P385	P385/1	55	Male	18	704.6585486	11.65082987	16.91431172	GE_ModelA	1.5	Tesla	GE Healthcare
P385	P385/2	55	Male	18	841.2006824	4.783474292	25.95964185	GE_ModelA	1.5	Tesla	GE Healthcare
P386	P386/1	42	Male	20	695.2950401	9.064204826	21.04966847	Philips_ModelA	1.5	Tesla	Philips
P387	P387/1	47	Female	18	667.6842195	8.740495921	37.121759	Philips_ModelA	1.5	Tesla	Philips
P388	P388/1	42	Female	29	775.2139212	5.002814885	30.69624279	Philips_ModelA	1.5	Tesla	Philips
P389	P389/1	52	Female	38	720.3896248	5.004032668	42.48707662	Philips_ModelA	1.5	Tesla	Philips
P390	P390/1	62	Female	21	710.3076061	9.101896488	15.94040554	Philips_ModelA	1.5	Tesla	Philips
P391	P391/1	25	Female	25	594.4356493	5.283801847	29.13007831	Philips_ModelA	1.5	Tesla	Philips
P392	P392/1	55	Not Appl	26	720.6275813	11.99598513	21.00406831	Philips_ModelA	1.5	Tesla	Philips
P393	P393/1	61	Male	24	658.8377661	9.469565342	29.0535058	Philips_ModelA	1.5	Tesla	Philips
P394	P394/1	45	Male	23	706.8308906	13.34468592	34.98758918	Philips_ModelA	1.5	Tesla	Philips
P395	P395/1	35	Male	16	820.261281	3.966198341	24.94297959	Philips_ModelA	1.5	Tesla	Philips
P400	P400/1	41	Male	30	674.6951289	7.951656637	25.32080813	Philips_ModelA	1.5	Tesla	Philips
P401	P401/1	39	Male	31	798.1599677	3.922555167	36.91744283	GE_ModelA	1.5	Tesla	GE Healthcare
P402	P402/1	33	Male	26	837.6224833	7.061598086	32.65986605	GE_ModelA	1.5	Tesla	GE Healthcare
P403	P403/1	42	Male	24	795.9273411	not quantifiable	35.4026815	GE_ModelA	1.5	Tesla	GE Healthcare
P404	P404/1	54	Male	28	739.4538584	7.293975656	31.65765981	GE_ModelA	1.5	Tesla	GE Healthcare
P412	P412/1	63	Female	30	812.47311	12.533321	30.03382	GE_ModelA	1.5	Tesla	GE Healthcare

TEST COMPARISON

Patient_ID	Scan visit	Age	Sex	BMI	liver_cT1	liver_PDFF	liver_T2star	Scanner
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	mismatch	match	match	match	match	match
match	mismatch	match	mismatch	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	mismatch	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match
match	mismatch	match	match	match	match	match	match	match

Appendix I Output Query Test Results

Females with age above 40 and BMI above 25 that have liver cT1 above 800 ms

Original Test Input Data Expected Results							
Patient_ID	Scan visit	Age	Gender	BMI	liver_cT1	liver_PDFF	liver_T2star
P206	1	60	Female	28	839.155028949975	1.197009769	24.11418677
P412	1	63	Female	30	812.473110000000	12.533321	30.03382

Test Results					
Patient	Visit	Age	Sex	BMI	livercT1
P206	P206/1	60	Female	28	839.155028900000
P412	P412/1	63	Female	30	812.473110000000

compare					
Patient	Visit	Age	Sex	BMI	livercT1
match	mismatch	match	match	match	0.000000050
match	mismatch	match	match	match	match

Siemens 1.5 Tesla visits (patients scans) where PDFF is below 5%

Original Test Input Data Expected Results								
Patient_ID	Scan visit	Age	Gender	BMI	liver_PDFF	Scanner	liver_cT1	liver_T2star
P106	1	36	Female	21	3.5693305104	Siemens_ModelA	865.1961088	31.52438818
P107	2	48	Female	26	2.3942956831	Siemens_ModelA	723.4435961	17.93460507
P108	1	48	Male	23	4.7287440318	Siemens_ModelA	812.696841	23.99313931
P108	2	48	Male	23	1.7789210516	Siemens_ModelA	871.4954244	21.38348531
P110	1	59	Female	28	4.2622986517	Siemens_ModelA	706.7094811	27.96504055
P110	2	59	Female	28	0.6349051208	Siemens_ModelA	597.1664494	45.10725443
P111	1	35	Female	32	4.1942506470	Siemens_ModelA	616.3865974	26.64438535
P112	1	45	Female	18	2.0207706022	Siemens_ModelA	708.3208524	17.379776
P113	2	45	Male	38	1.6506596133	Siemens_ModelA	571.0075321	43.87299637
P114	2	19	Male	22	3.7234793032	Siemens_ModelA	694.2238754	36.00094185
P115	1	50	Male	19	0.5351212102	Siemens_ModelA	692.34958	21.07063935

Test Results									
Patient	Visit	Age	Sex	BMI	liverPDFF	Scanner	Field Stre	Units	Manufacturer
P106	P106/1	36	Female	21	3.56933051	Siemens_ModelA	1.5	Tesla	Siemens
P107	P107/2	48	Female	26	2.394295683	Siemens_ModelA	1.5	Tesla	Siemens
P108	P108/1	48	Male	23	4.728744032	Siemens_ModelA	1.5	Tesla	Siemens
P108	P108/2	48	Male	23	1.778921052	Siemens_ModelA	1.5	Tesla	Siemens
P110	P110/1	59	Female	28	4.262298652	Siemens_ModelA	1.5	Tesla	Siemens
P110	P110/2	59	Female	28	0.634905121	Siemens_ModelA	1.5	Tesla	Siemens
P111	P111/1	35	Female	32	4.194250647	Siemens_ModelA	1.5	Tesla	Siemens
P112	P112/1	45	Female	18	2.020770602	Siemens_ModelA	1.5	Tesla	Siemens
P113	P113/2	45	Male	38	1.650659613	Siemens_ModelA	1.5	Tesla	Siemens
P114	P114/2	19	Male	22	3.723479303	Siemens_ModelA	1.5	Tesla	Siemens
P115	P115/1	50	Male	19	0.53512121	Siemens_ModelA	1.5	Tesla	Siemens

compare						
Patient	Visit	Age	Sex	BMI	liverPDFF	Scanner
match	mismatch	match	match	match	0.0000000004	match
match	mismatch	match	match	match	0.0000000001	match
match	mismatch	match	match	match	-0.0000000002	match
match	mismatch	match	match	match	-0.0000000004	match
match	mismatch	match	match	match	-0.0000000003	match
match	mismatch	match	match	match	-0.0000000002	match
match	mismatch	match	match	match	0.0000000000	match
match	mismatch	match	match	match	0.0000000002	match
match	mismatch	match	match	match	0.0000000003	match
match	mismatch	match	match	match	0.0000000002	match
match	mismatch	match	match	match	0.0000000002	match
	formatting				*decimal place error*	

cT1 is above 800 ms but PDFF is below 10%

Original Test Input Data Expected Results								
Patient_ID	Scan visit	Age	Gender	BMI	liver_cT1	liver_PDFF	liver_T2star	Scanner
P106	1	36	Female	21	865.19611	3.5693305	31.5243882	Siemens_ModelA
P108	1	48	Male	23	812.69684	4.728744	23.9931393	Siemens_ModelA
P108	2	48	Male	23	871.49542	1.7789211	21.3834853	Siemens_ModelA
P141	2	41	Male	22	822.57855	5.3744153	43.8386449	Siemens_ModelB
P152	2	39	Male	32	815.36532	5.5059484	29.06105	Siemens_ModelB
P158	1	44	Female	25	833.32157	4.3013299	34.2771293	Siemens_ModelB
P206	1	60	Female	28	839.15503	1.1970098	24.1141868	Siemens_ModelB
P212	2	60	Male	23	858.85676	9.8186471	35.8031235	GE_ModelA
P217	2	41	Female	23	946.74737	7.9830649	28.626317	GE_ModelA
P385	2	55	Male	18	841.20068	4.7834743	25.9596419	GE_ModelA
P395	1	35	Male	16	820.26128	3.9661983	24.9429796	Philips_ModelA
P402	1	33	Male	26	837.62248	7.0615981	32.659866	GE_ModelA

Test Results						
Patient	Visit	Age	Sex	BMI	livercT1	liverPDFF
P106	P106/1	36	Female	21	865.1961	3.569331
P108	P108/1	48	Male	23	812.6968	4.728744
P108	P108/2	48	Male	23	871.4954	1.778921
P141	P141/2	41	Male	22	822.5785	5.374415
P152	P152/2	39	Male	32	815.3653	5.505948
P158	P158/1	44	Female	25	833.3216	4.30133
P206	P206/1	60	Female	28	839.155	1.19701
P212	P212/2	60	Male	23	858.8568	9.818647
P217	P217/2	41	Female	23	946.7474	7.983065
P385	P385/2	55	Male	18	841.2007	4.783474
P395	P395/1	35	Male	16	820.2613	3.966198
P402	P402/1	33	Male	26	837.6225	7.061598

Compare Results						
Visit	**Patient**	**Age**	**Sex**	**BMI**	**livercT1**	**liverPDFF**
match	mismatch	match	match	match	-0.0000000089	0.0000000004
match	mismatch	match	match	match	0.0000000018	-0.0000000002
match	mismatch	match	match	match	-0.0000000161	-0.0000000004
match	mismatch	match	match	match	-0.0000000204	0.0000000000
match	mismatch	match	match	match	-0.0000000270	0.0000000001
match	mismatch	match	match	match	-0.0000000200	0.0000000003
match	mismatch	match	match	match	0.0000000500	-0.0000000002
match	mismatch	match	match	match	0.0000000411	-0.0000000003
match	mismatch	match	match	match	-0.0000000404	-0.0000000001
match	mismatch	match	match	match	-0.0000000434	0.0000000004
match	mismatch	match	match	match	-0.0000000118	0.0000000000
match	mismatch	match	match	match	0.0000000338	-0.0000000004
	formatting				*decimal place error*	*decimal place error*