Master's Thesis

Configurable Schema-Aware RDF Data Input Forms

Dávid Konkoly

April 2017

Albert-Ludwigs Universität Freiburg
Department of Computer Science
Chair of Databases and Information Systems

Candidate
Dávid Konkoly
Matr. number
3757311
Working period
18. 10. 2016 – 18. 04. 2017
Examiner
Prof. Dr. Georg Lausen
Supervisor
Victor Anthony Arrascue Ayala

Abstract

Abstract in English

Kurzfassung

Kurzfassung auf Deutsch

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Introduction.
You can reference the only entry in the .bib file like this: [2]

1.1 Initial goal and contributions

1.2 Thesis outline

Chapter 2

Preliminaries

2.1 Semantic Web

2.1.1 RDF

In RDF, abbreviation for Resource Description Framework, the information of the web is represented by means of triples. Each triple consists of a subject, predicate and object. The set of triples constitute to an RDF graph, where the subject and object of the triples are the nodes, the predicates are the edges of the graph. An RDF triple is called as well statement, which asserts that there is a relationship defined by the predicate, between subject and the object. The subjects and the objects are RDF resources. A resource can be either an IRI (Internationalized Resource Identifier) or a literal or a blank node (discussed later). A resource represents any physical or abstract entity, while literals hold data values like string, integer or datum. Basically there are two types of triples, the one that links two entities to each other, and the other that links a literal to an entity. The former expresses a relationship between two entities, and the latter in turn assign an attribute to the entity. Common practice is to represent IRI with the notation prefix:suffix, where the prefix represents the namespace, and the expression means the concatenation of the namespace denoted by the prefix, with the suffix. This convention makes the RDF document more readable. The namespace of RDF is the http://www.w3.org/1999/02/22-rdf-syntax-ns#, whose prefix is in most cases "rdf". This is defined on the following way:

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

Literals are strings consisting of two elements. The first is the lexical form, which is the actual value, and the second is the data type IRI. RDF uses the data types from XML schema. The prefix (commonly xsd) is the following :

@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

So a literal value in RDF looks as follows:

"Some literal value"^^xsd:string

The RDF vocabulary provides some built-in IRIs. The two most important are, the rdf:type property, and the rdf:Property class. The meaning of the triples, where the predicate is the property rdf:type is that the subject IRI is the instance of the class denoted by the object. Therefore the following statement holds in the RDF vocabulary:

rdf:type rdf:type rdf:Property.

It is maybe confusing that an IRI appears in a triple as subject and predicate as well, but we will see by the RDFS vocabulary that it is inevitable to express rules of the language. To be able to represent information about a certain domain, it is necessary to extend the RDF vocabulary with properties and classes. The classes will be discussed in the next section, but here it is explained how custom properties can be defined. The namespace of the example is the following:

@prefix eg: <http://example.org#>.

The example dataset intends to express information about people, which university they attend and how old are them. To achieve this two properties are needed:

eq:attends rdf:type rdf:Property .
eq:age rdf:type rdf:Property .

The actual data about a person:

eg:JanKlein eq:attends eq:UniversityOfFreiburg .
eg:JanKlein eq:age "21"^^xsd:integer .

2.1.2 RDF Schema

The previous section gave an insight into RDF world by showing how can information stored by means of triples. However the explanation did not mention that each RDF dataset has to have scheme, which is also called ontology. The ontology describes the set of properties and classes and how are they are related to each other. RDFS provides a mechanism to define such ontologies using RDF triples. The most important elements of the RDFS vocabulary can be seen on the following image.

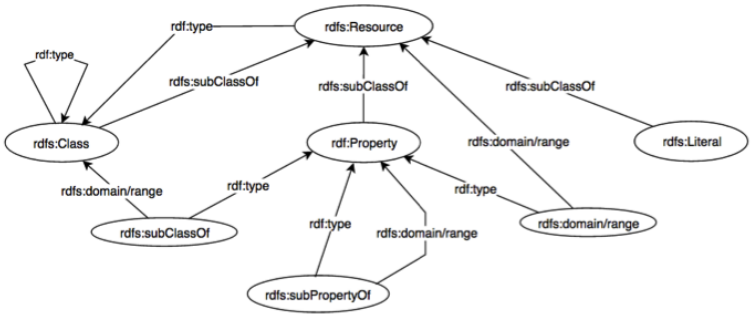

Figure 2.1: Main structure of the RDFS vocabulary

The two most important classes in the RDFS vocabulary is the rdfs:Class and the rdfs:Resource. The rdfs:Class is class, because it is the instance of itself, and the same way the rdfs:Resource is a class. The rdf:Propery and the rdfs:Literal are both classes as well. The rdfs:domain, rdfs:range, rdfs:subPropertyOf and rdfs:subClassOf are properties. Important to note that these properties are subjects and predicates in the same time in the RDFS vocabulary graph. Also they describe themselves like rdf:type. The properties rdfs:domain and rdfs:range describe for the property the type of the subject and object respectively, which with it can build a triple as predicate. The following image illustrates their meaning:

Since both the rdfs:domain and rdfs:range are properties themselves, they have as well their domain and range, which is the class rdfs:Resource. The property rdfs:subClassOf expresses subclass relationship between classes. It means if a class B is a subclass of class A, and resource R is the instance of class B, then resource R is the instance of class A as well. Since it describes the relationship between two classes its both domain and range is

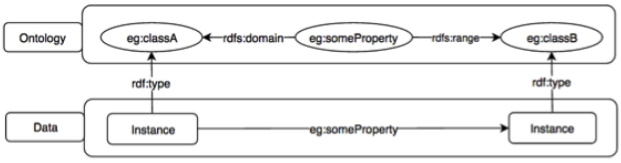

Figure 2.2: RDFS domain and range definition

the class rdfs:Class. The property rdf:subPropertyOf expresses the relationship between two properties. If property P2 is sub property of P1 and two instances are related by P2 then they are related by P1 as well. Its domain and range is the class rdf:Property. Now everything is given to define the ontology for the example of the previous section.

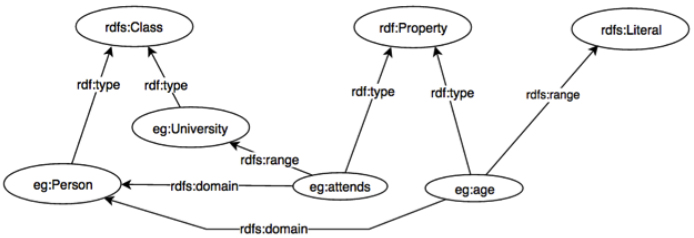

Figure 2.3: RDFS domain and range definition

2.1.3 OWL

OWL, abbreviation for Ontology Web Language is an extension of the RDFS vocabulary. OWL allows expressing additional constraints on the data, above the range and domain definitions. These constraints are called restrictions. Restrictions are conventionally expressed by blank nodes. Blank nodes do not have IRIs, but it is defined through the triples in which they participate as a subject. For example a restriction stating that the instances of the class eg:FootballTeam can build a triple through the eg:hasPlayer property only with the instances of eg:FootballPlayer class can be expressed the following way:

```
eg:FootballTeam rdfs:subClassOf [
  rdf:type owl:Restriction ;
  owl:onProperty eg:hasPlayer ;
```

```
  owl:allValuesFrom eg:FootballPlayer .
]
```

Listing 2.1: OWL restriction in N3 format

owl:Restriction is class and owl:onProperty and owl:allValuesFrom are properties. It can be seen that class, on which the restriction applies is the subclass of the restriction blank node. Furthermore OWL is capable of expressing qualified cardinality restriction. For example the statement that a basketball team has to have exactly five players, look as follows in OWL:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX eg: <http://example.org>

eg:BasketballTeam rdfs:subClassOf [
  rdf:type owl:Restriction ;
  owl:onProperty eg:hasPlayer ;
  owl:onClass eg:Player ;
  owl:qualifiedCardinality "5"^^xsd:nonnegativeInteger
] .
```

Listing 2.2: OWL restriction in N3 format

These two examples cover the thesis related features of OWL. The next image depicts the OWL vocabulary.

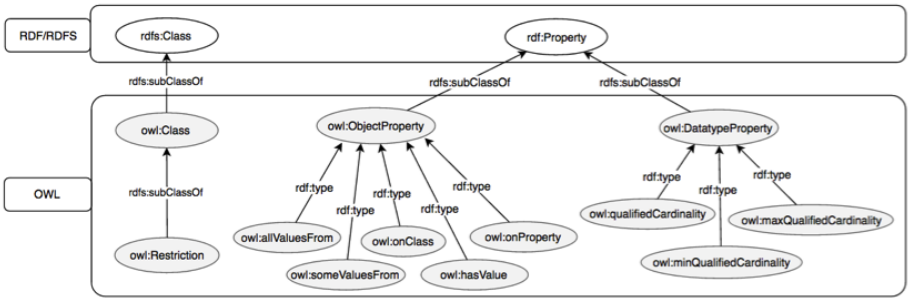

Figure 2.4: A subset of OWL vocabulary

There are two new class types are the owl:Class and the owl:Restriction. The rdf:Property has two subclasses, the owl:ObjectProperty and owl:DataTypeProperty. owl:ObjectProperty represent the properties that links instances to instances, and the owl:DataTypeProperty is those that link instances to literals. The

following two images shows the domain and range definitions of the OWL properties used to describe restrictions.

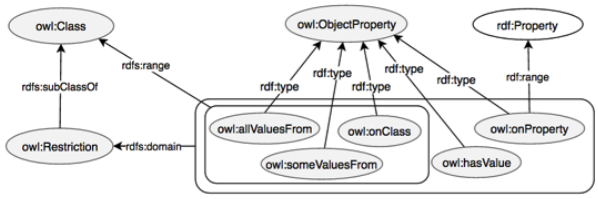

Figure 2.5: OWL object properties

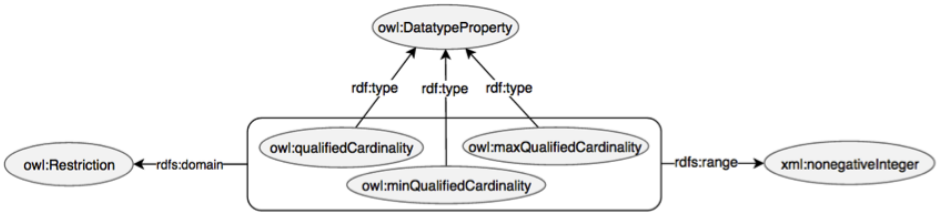

Figure 2.6: Properties for qualified cardinalities

2.1.4 SPARQL

SPARQL is a query language for querying data in RDF graphs. A SPARQL query is a definition of a graph pattern through variables and constants. The following example query returns all IRIs that represent a football player:

```
SELECT ?player
WHERE {
  ?player   rdf:type   eg:FootballPlayer .
}
```

Listing 2.3: SPARQL Query I.

In the example the query consist of only one triple. The subject is a variable and the predicate and the object are constant. Therefore the triple store looks all the triples and checks the predicate is rdf:type and the object is eg:FootballPlayer. It is well possible to not just ask the IRI of the players but further information by adding additional triples to the query in order to ask the name for example of the player:

```
SELECT ?player ?name
WHERE {
  ?player   rdf:type   eg:FootballPlayer .
  ?player   eg:name    ?name .
}
```

Listing 2.4: SPARQL Query II.

The result table in this case will contain two columns, one with the IRI of the person and one with their name. Important that it is as well possible to query blank nodes by introducing a variable for it. So if we want to list all the instances that are coming into question as player to a football team we can formulate the following query:

```
SELECT ?person ?name
WHERE {
  eg:FootballTeam rdfs:subClassOf   ?restriction .
  ?restriction    rdf:type          owl:Restriction .
  ?restriction    owl:onProperty    eg:hasPlayer .
  ?restriction    owl:allValuesFrom ?playerType .
  ?player         rdf:type          ?playerType .
  ?player         eg:name           ?name .
}
```

Listing 2.5: SPARQL Query III.

2.2 Applied Ontologies

Ontologies are used to describe types, relationships and properties of objects of a certain domain. It is a common practice to used already defined ontologies rather developing an own. The first reason is, that the development of an ontology is a complex and a tedious process, and requires a lot of resource. Secondly, it is reasonable to use standardized vocabularies, in order to make data from same domain but different sources inter-operable.

2.2.1 Foundational Model of Anatomy - *FMA*

The foundational Model of Anatomy ontology is an open source ontology written in OWL. FMA is a fundamental knowledge source for all biomedical domains, and it provides a declarative definition of concepts and relationships

of the human body for knowledge based applications. It contains more than 70 000 classes, and 168 different relationships, and organize its entities into a deep subclass tree [4]. All types of anatomical entities are represented in FMA, like molecules, cells, tissues, muscles and of course bones. In our project we use only the subset of the FMA. The taken elements are the subclasses of the following two classes and the three properties:

- Classes

  Subdivision of skeletal system - fma:85544

  Bone Organ – fma:5018

- Properties

  fma:systemic_part_of

  fma:constitutional_part_of

  fma:regional_part_of

The class *Bone Organ* is the superclass of all bones in the human skeleton. Each bone belong to a skeletal subdivision and a skeletal subdivision can be a part of another skeletal subdivision. This relationship in both cases is expressed by the property *fma:systemic_part_of*. To define which bone organ belongs to which skeletal subdivision FMA uses OWL restrictions (see Figure 2.7). The properties *fma:constitutional_part_of* and *fma:regional_part_of* *fma:constitutional_part_of* are discussed later.

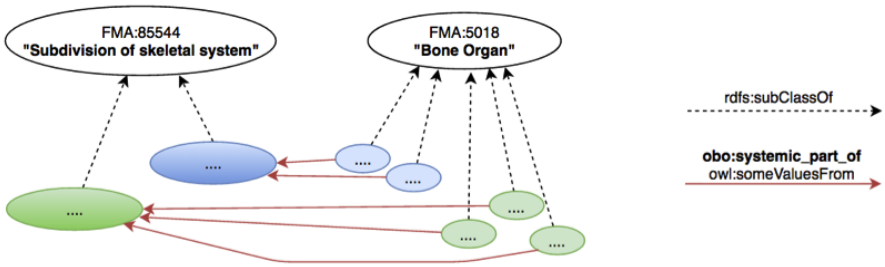

Figure 2.7: Ontology structure for skeleton

Finally the advantage of using the FMA ontology is that, if in the future further elements of the human body have to be addressed by the research

processes, i.e. muscles, then these classes can be easily integrated to the currently applied subset.

2.2.2 Ontology for Biomedical Investigations - *OBI*

The aim of OBI ontology, is to provide the formal representation of the biomedical investigation in order to standardize the processes among different research communities. It is a result of a collaborative effort of several working groups, and it continuously evolving as new research methods are being developed. Its main function to describe the rules how biological and medical investigations have to be performed. OBI reuses terms from BFO *Basic Formal Ontology* IAO *Information Artifact Ontology* and OBO *Open Biological and Biomedical Ontologies*[3]. To define processes OBI uses the following three general classes:

- *Information Content Entity* - obo:IAO_0000030
- *Material Entity* - obo:BFO_0000040
- *Process* - obo:BFO_0000015

Information Content Entity represent results of a specific measurement, while Material Entity stands for the objects, on which the meausurements have been performed. The Process could mean any kind of step within an investigation, from the planning, through execution till the conclusion.

- *Planning* - obo:OBI_0000339
- *Study Design Execution* - obo:OBI_0000471
- *Drawing a conclusion* - obo:OBI_0000338

In our project the following three properties are used:

- *has part* - obo:BFO_00000051
- *has specified input* - obo:OBI_00000293
- *has specified output* - obo:OBI_00000299

2.3 Web applications

This chapter contains practical information about how web applications work. In section 2.3.1 the basic mechanism of data driven applications are discussed, like navigation between page, data display and creation. Section **??** then focuses on the applications that are using semantic technologies, and addresses what kind of architectural changes that means.

2.3.1 Client-sever architecture

A web application is program that runs on a machine, which is accessible through the web. The machine is called server, because its main purpose is to server request that are coming from the web browser. Web browsers are as well programs, but they run on personal computers, tablets, etc, and they are capable of sending request through web to the servers. The response to these requests are HTML document, which can be displayed by the browser.

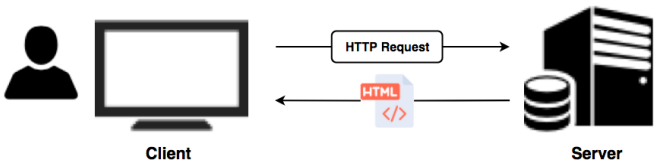

Figure 2.8: Client server communication

An HTML document contains definition of the elements of the pages, such as tables, buttons, etc. It contains as well so called CSS docuements (Cascading Style Sheet), which is responsible for the definition of the style of the elements. Moreover to make the web pages more interactive, JavaScript (JS) de can be embedded to HTML as well.

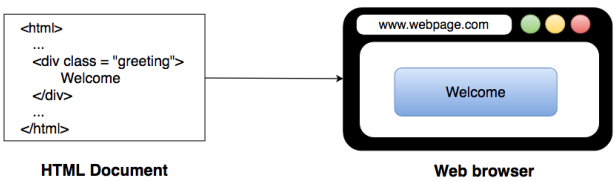

Figure 2.9: HTML document is interpreted by the browser

Initially web pages were static, which means that their only function was

to show certain set of information. These applications usually web applications do not consist of one single page, but of several different pages. Like a web page for news, have normally a main page, and different sub pages for the particular topics. In order to navigate between the pages of the application, the HTML document contains links that trigger further HTTP requests. Links in HTML can be defined by means of the <a/> tag. The most important parameter of this tag is href, whose value contains the URL of the HTTP request. Let assume that an application's main page is accessible through the URL http://newsPortal.com. Common practice that subpages of the application can be called through various url-mappings, which means the main URL is extended with a keyword that denotes the page to be requested.

```
<a href="http://newsPortal.com/politics"> Politics </a>
<a href="http://newsPortal.com/sport"> Sport </a>
```

Listing 2.6: Example link definitions

If the user clicks on of these link (with the label 'Politics' and 'Sport') then these request will be sent to the news portal page. Each such request has to be served, differently to each mapping some routine has to be assigned. For example by Java web applications, the classes of the server that process the request are called servlets. On the next image it is shown, how the XML file defines, which class is responsible for the the mapping '/politics'.

```
<servlet-mapping>
  <url-pattern>/politics</url-pattern>
  <servlet-class>servlets.PoliticsController</servlet-class>
</servlet-mapping>
```

Listing 2.7: Java servlet mapping definition

Then the responsibility of the class *servlets.PoliticsController* is to respond the corresponding HTML page for the client. Figure 2.10 show the main structure of the applications, where the rectangles on the client side represent the different pages of the application.

2.3.2 Data driven web applications

This section aims to present the fundamentals of the web technologies that allows to build application for browsing and creating data. Modern web ap-

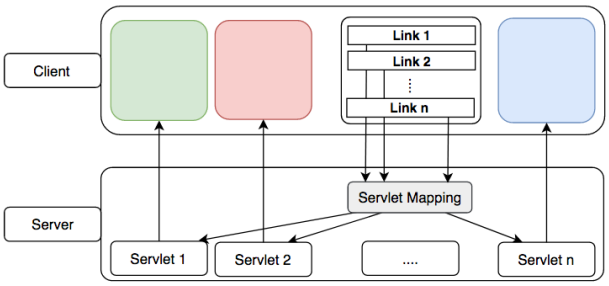

Figure 2.10: Navigation through the web application

plications do not store the information in HTML documents. So the page loading process is not just the sending the HTML document, but a retrieval of a particular dataset, and the substitution into a web page. First of all the task of responding requires a query that retrieves that data from the database. By applications using relational data model, the tables and attributes are always modeled by classes of the used object oriented programming (OOP) language. So the data retrieval is the instantiation of the classes in scope.

Let assume that articles of a news portal is stored in a table with the attributes, id, type, title, summary and text. Then there has to be a class defined in the server code with the same attributes. To instantiate instances of the class, it is necessary to perform an SQL query.

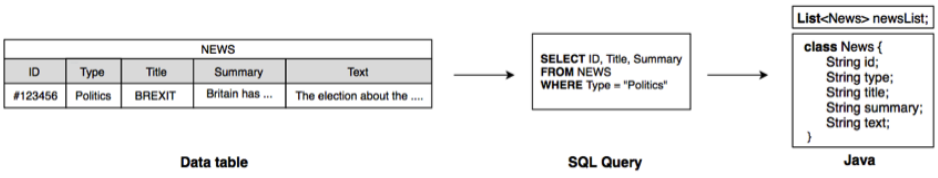

Figure 2.11: Data flow

The query results not only one instance of the News class, but a list (*List<News> newList*). To generate from this a HTML page that shows the articles, normally so called template engines are used. Templateing enables to define the HTML documents parametric, and passing them data, and they generate the result page automatically.

```
<#list  newsList as news>
```

```
  <h3>  ${news.title} </h3>
  <p>  ${news.summary} </p>
  <a  href = "http://newsPortal.com/wholeNews?id=${news.id}">
    Read more
  </a>
</#list>
```

Listing 2.8: Template file example

The template file is a description of how the data has to be converted into HTML document. It can be seen that it is possible for instance to declare a list on the input variable newsList. Then the template engine iterates through the News objects and by accessing its fields (title, summary, id) and generates the HTML for each element. So the complete flow of data from the database to the client looks as follows:

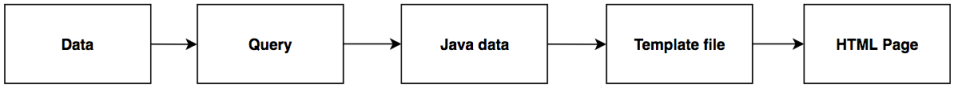

Figure 2.12: Flow of information from DB to client

The template shows only the summary of the article, but offers the following link:

```
http://newsPortal.com/wholeNews?id=${news.id}
```

The new feature is that after the url mapping there is a parameter *id*, and its value will be the database id of the web application. The idea is that this link redirects to the page where the whole article can be seen. So there has to be a servlet class defined to the mapping /wholeNews, which to perform the following query where the id is the input.

```
SELECT Text
FROM NEWS
WHERE ID = ${id}
```

Figure 2.13: SQL query with parameter

Thus it is achieved that different links are programmed to get access not only to different other pages, but to specific data items.

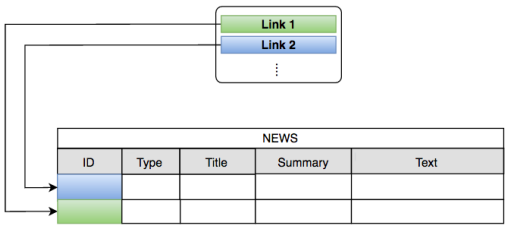

Figure 2.14: Links to data items

Web applications do not only just display existing data, but they allow the users to enter their new data. In HTML the element used for data input is called form. Form is a container, and it consists of particular form elements according to the data to be added.

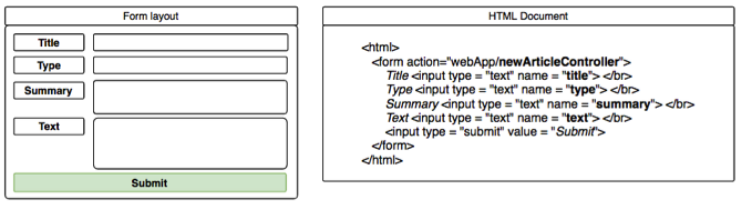

Figure 2.15: Form layout and HTML document

Submitting the form to the server send an HTTP request with multiple parameters, where they are divided through the & character.

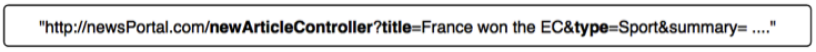

Figure 2.16: Request with parameters

By the data entry creation the task of the controller is to get the values from the request an instantiate the class representing the data to be created. Then initialized class instance is passed to the database where the entered data will be persistently stored.

2.3.3 Applications with RDF Data

This section aims to give an insight to web application that are based on RDF data. It will be covered what kind of requirements do the software have on the server side to create RDF data, and what is the difference between the

```
String title = request.getParameter("title");
....
News news = new News(title, type, summary, text);
DatabaseConnector.insert(news);
```

Figure 2.17: Example Java routine for data storage

RDF model based applications and the relational ones. The most important feature of RDF that the data scheme, namely the ontology is stored in RDF triples too, thus can be queried.

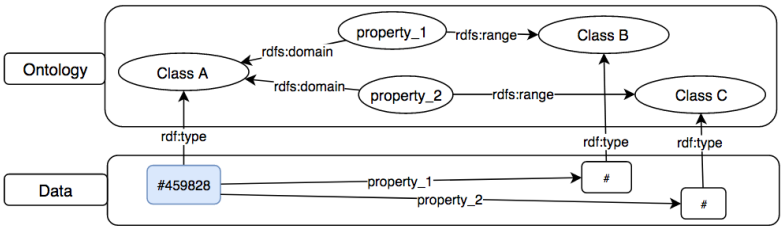

Figure 2.18: Ontology and data in RDF

Therefore it is possible to generate web pages that can adapt to the ontology. The following query demonstrates that how it is possible to get all the instances, which are connected to a particular instance (#459828).

```
SELECT ?property ?relatedInstance
WHERE {
  ?instance       rdf:type        ?class .
  ?property       rdfs:domain     ?class .
  ?instance       ?property       ?relatedInstance .
  FILTER (?instance = #459828 ) .
}
```

Listing 2.9: Dynamic SPARQL query

The first two lines of the query defines the properties whose domain class is the type of the input instance, while the third asks for all triples with the possible properties. If the result of the query is then grouped based on properties, then the dataset can be displayed by a template using two lists. The outer list ceates fields for the properties, and the inner show all the instances with that property.

```
<#list properties as property>
  <#list property.dataSet as instance>
```

```
  <#list>
</#list>
```

Listing 2.10: Ontology adaptive template file

The VIVO framework applied in the RDFBones project generates the pages for instances this way.

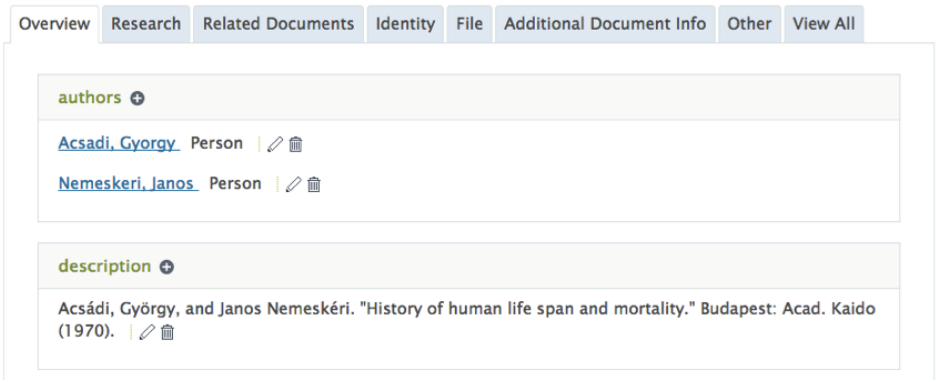

Figure 2.19: VIVO Profile page

Above the display of the existing data, RDF based applications are different as well in the data input mechanism. First of all, due to the fact that ontologies can contain thousand of class it is not an option to represent them all as classes of the server application language as well. It is time consuming and the system would loose its flexibility in the cases when new ontology subsets are supposed to be loaded. Therefore there are neither for each type of the database an entry form with a unique controller servlet, but more generic approaches are used. To define what dataset has to be created, semantic web based applications simply define them as a set of triples, with variables like in SPARQL, just the data flows the other way around.

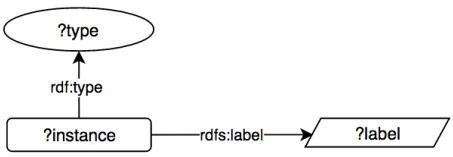

Figure 2.20: Triples representing a new instance

Figure 2.20 simple set of triples that have to be created by new data entry generation. The value of the variable textit?instance will be IRI that have

not been used. This is an essential part of every triple store implementation that they provide unused uris for the server application for the new set of triples. The values of the *type* and label, are coming from the input form. The label is just like in the previous example, it is a string typed by the user it is stored as an attribute of the new entity. But the type of the instance is class IRI. The point is that the options of the selector field, from which the type value is coming, is filled with the results of a SPARQL query on the ontology. So for example if the entry form provide the possiblity to create any type of processes from the OBI ontology then the before the form loading the following query has to be executed.

```
SELECT ?class ?label
WHERE {
  ?class       rdf:subClassOf     obo:OBI_0000339 .
  ?class       rdf:label          ?label .
}
```

Listing 2.11: SPARQL query for the input form

will be a new unused IRI of the triple store, while the ?type variable come from the client from a selector field. The following image depicts this simple scheme of the data input process.

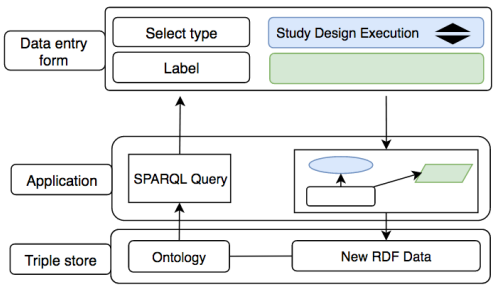

Figure 2.21: Data input scheme by RDF

Chapter 3

Problem Statement

This chapter is divided into three sections. As the application is highly dependent on the underlying data model, the first section is dedicated to the data scheme describing the investigations in scope. The second chapter in turn addresses the problem of web applications that allows the creation of RDF data explained in the first section. It covers the issues of both the client and server side implementation and their communication. Finally section 3.3 outlines the scheme of the solution proposed and implemented by the thesis work.

3.1 Modeling anthropological research activity

This section consist of three subsections. The first two (3.1.1 and 3.1.2) describes how the RDFBones ontology (developed during the project) integrates the *FMA* and *OBI* ontologies for describing research processes related to anthropology. While the third section (3.1.3) discusses how can the core ontology be extended to define custom bone segments and processes.

3.1.1 Data on skeletal remains

We have seen in section 2.2.1 the base structure of the human skeleton. The most important point is that not only individual bones will be represented in the data we create, but the skeletal regions as well, like skull or vertebral column. The institute where these investigations are conducted posses mainly skeletal remains of skulls. The skull has the peculiarity that it does

not consists directly of bones organs, but from two sub skeletal divisions, and these two subdivisions contain the bone organs. Figure 3.1 shows the ontology subset for the skull and the data instances (each denoted with #). The red arrows denotes restrictions on the properties *fma:systemic_part_of*.

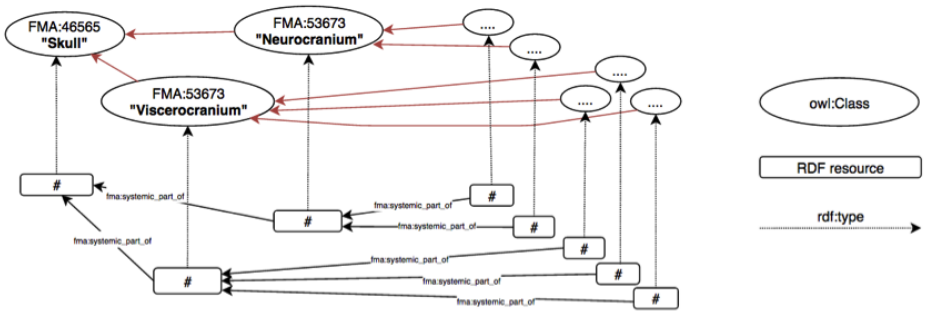

Figure 3.1: Ontology and triples of the skull

So we know how skull and its subdivisions and bone organs are represented by RDF, but there are processes where specific bone segments have to be addressed as well. Therefor RDFBones have the class *rdfbones:SegmentOfSkeletalElement*. This instances of this class is connected to the instances of the class *Bone Organ* with the property *fma:regional_part_of*.

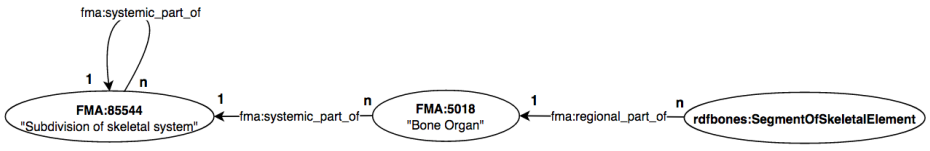

Figure 3.2: Bone segment in RDFBones

Furthermore these instances representing skeletal remains do not stay in the database individually. If a researcher takes a specific skull from the collection of the institute, it makes a so-called skeletal inventory, which records what bones segments are complete, partly present or missing. To store these information in the RDFBones ontology introduces three further classes (above the one for bone segment), which are all the subclasses of *OBI* classes.

Figure 3.4 illustrates the dataset through a simple problem. The upper left part of the figure shows that a specific bone is divided into three bone

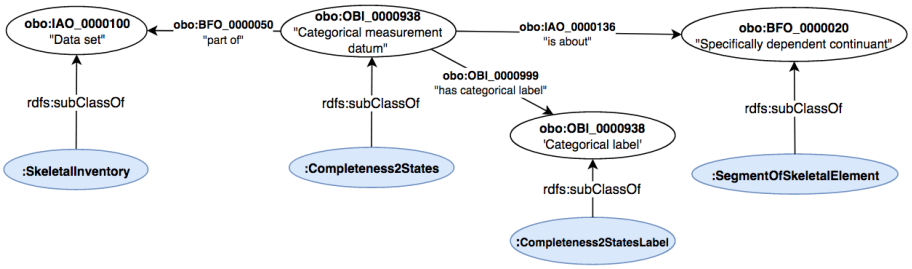

Figure 3.3: RDFBones as extension of OBI

segments, while the right shows the existing bone from which the data has to be stored. It can be seen the section I. is complete, section II. is just partly present and the III. is missing. The lower part of the figure then shows that the three segment are represented as subclass of the *:SegmentOfSkeletalElement* class (denoted with blue), and from the third type there are no instance have been created, while the first two are connected to the skeletal inventory and compleness label instances through *:Completeness2States* instances.

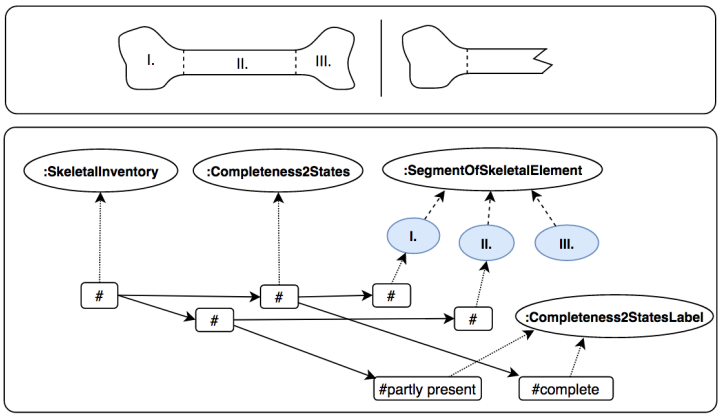

Figure 3.4: Custom bone segment example

This is the way how the information is stored about skeletal remains using the RDFBones ontology.

3.1.2 Investigation process

An investigation is done by execution of the a study design. It has three parts, the assay, the data transformation and the drawing of a conclusion.

From assays and data transformations there can be more in one execution, but there is only one conclusion. The inputs of the assays are always segments of skeletal elements, an their output is always a measurement datum, while the data transformation's input and output are both measurement datums.

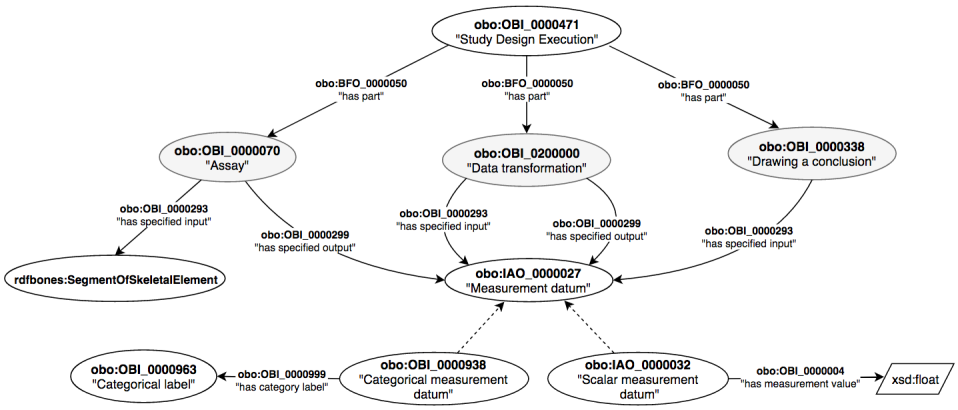

Figure 3.5: Applied subset of OBI ontology

To understand a bit more what this data model can be actually used, let us take the example of an investigation, whose goal is to determine if a taken skull belonged to a male or female. The basis is that the male and female skeleton has different peculiarities that can be quantified, how expressed they are. Figure 3.6 illustrates an example the token *Glabella*, which is on the *Nasal bone*, and its expressions.

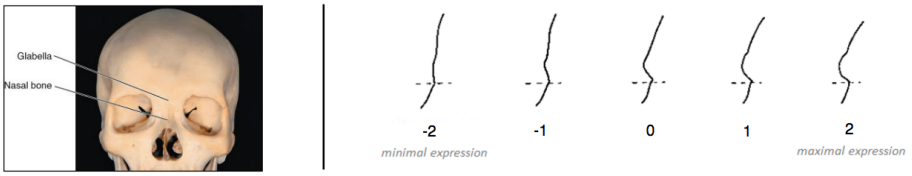

Figure 3.6: Glabella and its expressions

The larger the numbers for masculine and the lowers are feminine expression. An assay in this case is an assignment of a scalar value to a bone segment. The investigation process does not take only one bone segment but several different ones, to reduce the possibility of the erroneous output. Fig-

ure 3.7 shows a dataset of a study design execution where the green arrows stand for *has specified input* and the blue ones for the *has specified output* predicates. The idea is simple, the output of the assays are aggregated, and if the output is smaller then zero then it was a male, otherwise a female.

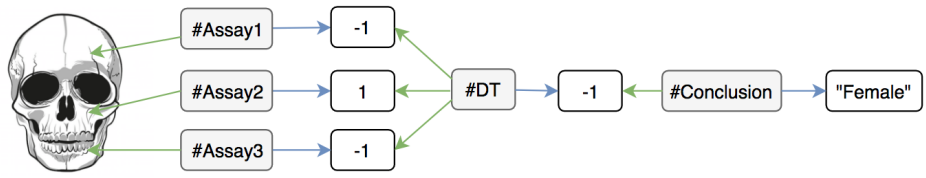

Figure 3.7: Study design execution dataset

Where the light grey boxes are the instances of *Assay*, *Data Transformation* and *Drawing conclusion* classes respectively.

3.1.3 Ontology Extensions

The previous two sections introduced the data scheme of the problems of the project. This part provides a more detailed explanation about how exactly these ontologies can be extended to tackle custom problems. By skeletal inventories *RDFBones* ontology allows to define custom bone segments of the bone organs in order to enable more fine-grained representation of the research activity. However by most of the cases it is sufficient to address the bone as a whole. For such cases *RDFBones* has the class *:PrimarySkeletalInventory* which encompasses all the entire bone organs. The definition of this inventory can be seen on the upper part of Figure 3.8.

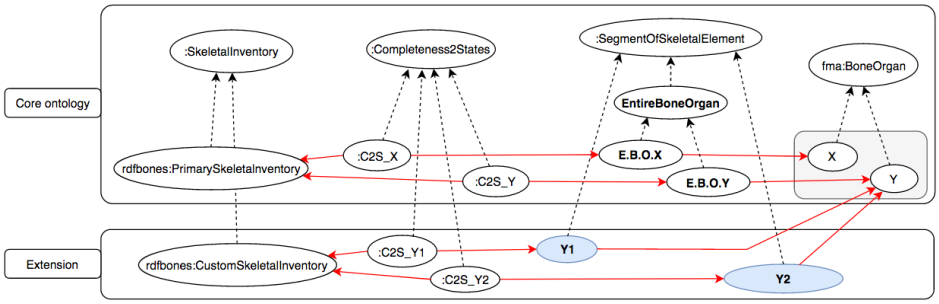

Figure 3.8: Ontology extension for skeletal inventories

The class *:EntireBoneOrgan* has as many subclass as many bone organs are there. The classes X and Y represents the set of all bone organs. Furthermore each of the entire bones are connected to the custom subclasses of *:Completeness2States*, to establish the connection to the primary skeletal inventory. The reason why there is not only one entire bone class for all bone organs, is that the input of the assays must be a bone segment, and it must be possible to address each of them individually. The lower part of Figure 3.8 shows a custom inventory definition, by new completeness and bone segment classes.

By the study design execution only the assay part will be covered, because the data transformation has exactly the same structure. As it was already mentioned through the example from Figure 3.7, each assay must address a bone segment an input, and a measurement datum as an output. To define a custom study design execution (Figure 3.9) at first a subclass is needed (:SexExstimation), and the different custom assays (denoted with blue) that are connected to each other again with restrictions. Then the outputs of the assays are custom scalar measurement datum classes.

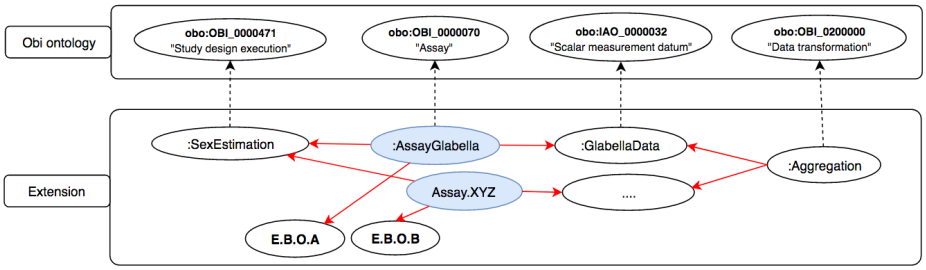

Figure 3.9: Ontology extension for sex estimation

3.2 RDF Data input

3.2.1 Dynamic data entry forms

In section **??** it was explained that the simplest data input process by RDF data is the substitution of the values coming from the client into a predefined set of RDF triples. In this case the interface is a static HTML form, the input data is a set of key-value pairs, while the RDF triples are defined on the server

simply by a string. However in the previous section we have seen that the data models of the problems contain relationships with 1 to n cardinalities. This means that the data input process may contain, above a selection of the type, and setting some of the literal values of one new RDF instance, the dynamic adding of sub forms that represent further instances that are connected to the main instance with cardinality n.

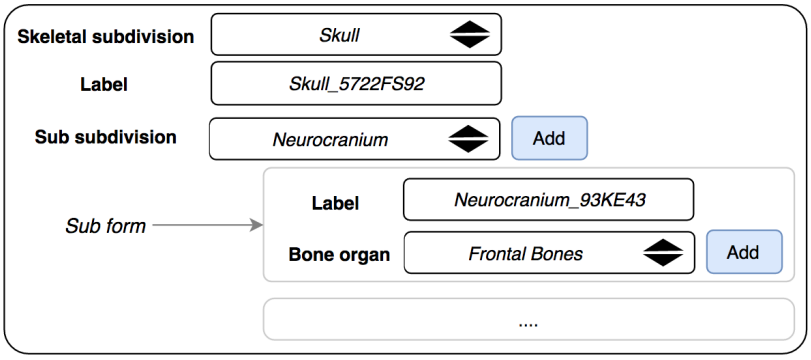

Figure 3.10: Multi dimensional form layout

Figure 3.10 depicts a form which allows to add new sub forms that represent the sud subdivisions of the selected skeletal subdivision. Such functionality can be implemented by means of JavaScript routines embedded into the HTML document. The idea is that there is a main form, which consist of the same type of elements like the static forms, and one of its selector field is equipped with a button that initiates the loading of a new sub form. Then the sub form consists of again the standard elements, and maybe of further sub form adders. In the example the sub sud subdivision form contains an other sub form adder field for the bone organs.

The task of the JavaScript code running on the client is handle these click events and add the new form elements to the HTML document dynamically. Moreover it has to access the input elements and insert the values into the form data object with defined keys. As there can be multiple sub forms added to the forms their data object stored by means of arrays. Listing 3.1 shows the JSON data (JavaScript Object Notation) of the form in Figure 3.10.

```
var formData = {
  subdivisionType : "fma:5018",
  subDivisionLabel : "Skull_5733FS2",
  systemic_parts : [
```

```
    {
      subsubdivisionType : "fma:45720",
      subSubdivisionLabel : "Neurocranium:_93KE43",
    }, { ... } ]
}
```

Listing 3.1: JSON object generated by the form

The server then in turn has to be prepared that a particular subgraph of the data model has to be created multiple times, and the values are arriving in arrays. The following image shows the data model of the entry form.

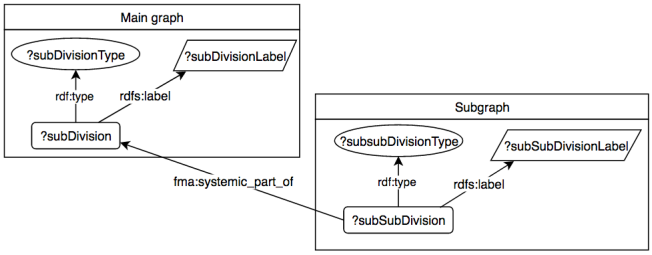

Figure 3.11: Multi dimensional RDF dataset

3.2.2 Adoption form elements to the ontology

The previous section showed the basics of forms for multi dimensional data input. An important issue which was just partly addressed in 2.3.3, is how the options of the particular selector fields are loaded. In order to load the options of the initial selector can template files be used, that gets the list of the classes coming from a SPARQL query on the ontology. In the example these are all skeletal subdivisions. But there are elements, for example the second selector field in Figure **??**, whose values are dependent on the previous elements, which his means their options must be loaded dynamically only after selector of its ancestor element.

Therefore their values are not part of the initial HTML document like possibly in the case of the independent skeletal subdivision class selector, but they have to be set as values of JS variables, so that they can be loaded after the selection. This can be achieved the same way by template file routines, that gets the set of sub sudivisions grouped by skeletal division.

```
var subSubDivisions = {
```

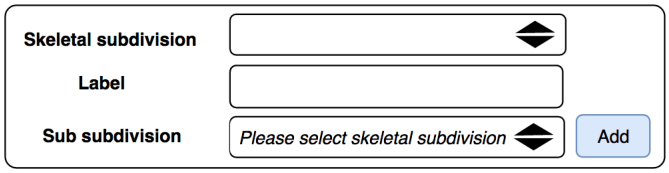

Figure 3.12: Initial state of the form

```
  skull : [{ neurocranium } ... { .. }],
  vertebralColumn : [{ cervical vertebra } ... { .. }],
  ...
}
```

Listing 3.2: Form option data structure in JSON

The data set on the listing 3.2 stores the lists of the sub ubdivision with the corresponding keys (the keys are real IRIs in practice). Such data can be written to JS with template routines by means of two lists, like in listing 2.10. Then if the user selects the first field, JS can access based on the selected value the array to be loaded as options.

But as the example on Figure Figure 3.10 shows that after adding the sub subdivision the bone organs are supposed to be selected to. Therefore the client has to have the same data structure for bone organs as well grouped by the sub sudivsion. The problem is that this data on the form can easily became too large, which may lead to performance issues. Additionally if the whole dataset is queried by the page load then it can be as well inefficient. Therefore it is common a practice that data of the form is loading dynamically through AJAX (Asychronous JavaScript and XML). AJAX is technology that allows the client to communicate with the server without reloading the whole page. This makes the pages more interactive and as well more efficient.

The idea is that JS is programmed so that upon certain events, a particular request data is assembled and sent to the server. The response then in this case is not a whole HTML page but a dataset in XML or JSON format. After the response arrives, a JS routine continues its. This more elegant solution addresses further JS programming, and of course new server routines, that handles the requests, perform the queries with the incoming parameters, and respond the results.

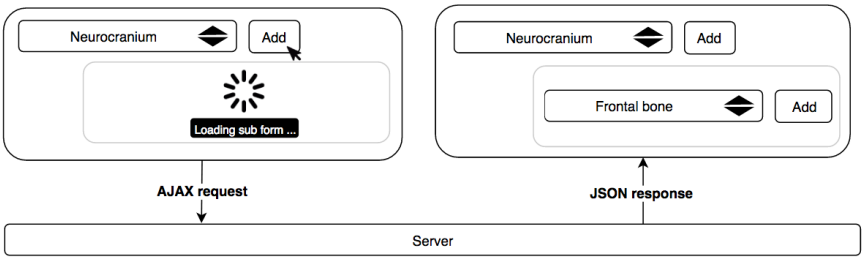

Figure 3.13: Loading form data through AJAX

3.2.3 Editing form data

The dataset created by the forms have be browsed and edited as well. This means that the form has to be restored to the same state as it submitted for initial data creation. This is the other direction of the data flow. Here the existing data will be retrieved by SPARQL, JSON object is generated and the form algorithm generates the form by adding the sub forms based on the data.

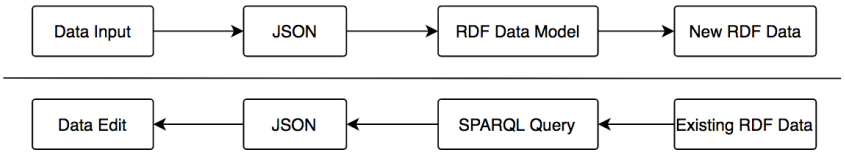

Figure 3.14: Two directions of the data flow

Important issue regarding the server implementation is that in such multi dimensional dataset, it is not sufficient to perform only one query for the whole form data. For example there can be more elaborate assays that have multiple inputs and output (i.e Figure 3.15).

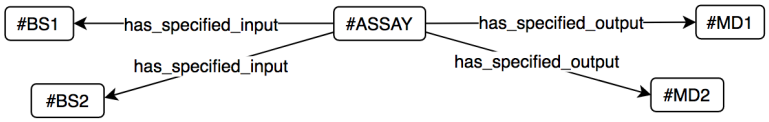

Figure 3.15: Multi dimensional data example

This means if the following query is executed on the dataset, then the result is such a table ((Table 3.1)) that is hard to process.

?boneSegment	?measurementDatum
BS1	MD1
BS1	MD2
BS2	MD1
BS2	MD2

Table 3.1: SPARQL result table

```
SELECT ?boneSegment ?measurementDatum
WHERE {
  ?assay     :has_specified_input         ?boneSegment .
  ?assay     :has_specified_output        ?measurementDatum .
  FILTER ( ?assay = ${inputParameter})
}
```

Listing 3.3: SPARQL query for the form data

Therefore the data object of the form has to be retrieved gradually, by dividing the data model graph by the predicates, whose cardinality is larger than one. The process have to start with main graph, and the results of the sub graph queries then has to be stored arrays of the form data JSON object. Based on this data the task of the client is interpret the data and generate again the form. Important part of the form reloading based on existing data is that, not only the values of the literal fields have to be restored but the selector field options as well. So that for example if the existing dataset contains a skeletal subdivision, it is necessary to load to the subform the possible bone organs into the selector, so that the user can add additional bone segments conveniently. Finally if a value of a literal field changes, or new sub forms has to be added or removed, the entry form data should not be completely sent again to the server, but only the data fields that are concerned by the modification. Thus it does not require a complete page reload, and these operation can be performed again through AJAX calls. To achieve this the client has to be prepared to be able to send data modification requests to the server on change event of any form element or sub form.

3.2.4 Saving data

3.3 Solution Scheme

Chapter 4

Implementation

4.1 Web application ontology

4.1.1 Introduction

- Computer programs are able to overtake exercise from humans to accelerate work. In this case work we want speed up is the application development.

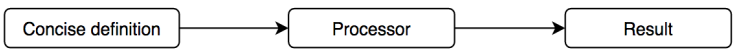

Figure 4.1: Basic workflow

- For example we have the form definition.

```
<form action="http://example.org/newUser">
    Username <input type="text" name="userName"></br>
  Password <input type="text" name="password"></br>
  <input type="submit">Submit</input>
</form>
```

- Can be modelled by a configuration dataset

```
{
  type : "form",
```

```
  action : "http://example.org/newUser",
  elements : [
    { type : "text", text : "Username", varName : "userName"},
    { type : "text", text : "Password", varName : "password"}
  ]
}
```

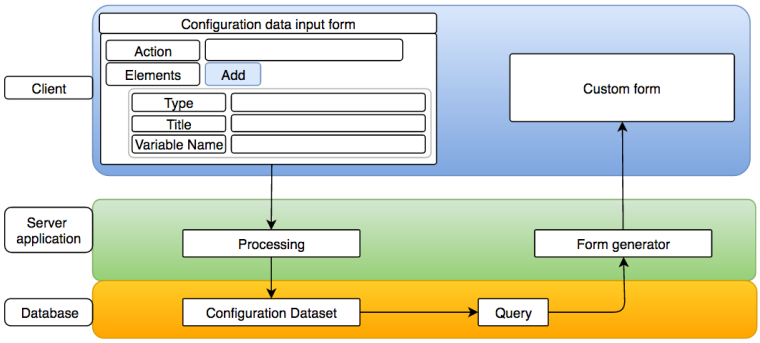

Figure 4.2: Framework functionality

- The goal is store the configuration of the web application persistently in a database. In our case the goal is to store the data in RDF data.

4.1.2 Form definition

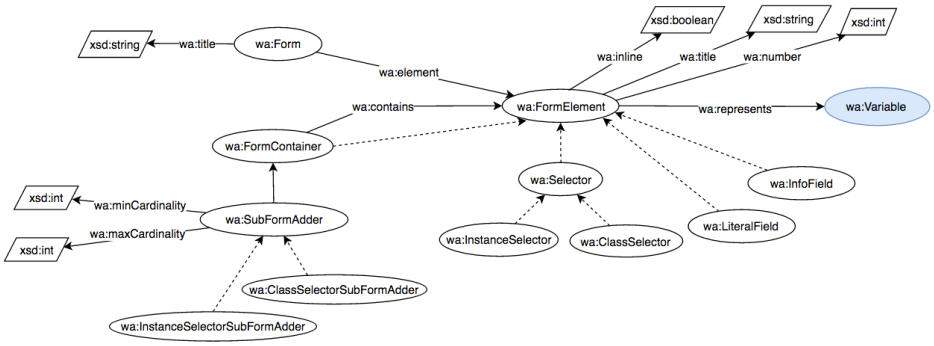

Figure 4.3: Web application ontology for the form

- Explanation of the main elements and their functionality

- Next that each form element has to represent a variable. This is denoted with blue the wa:Variable. And the variable name from Image X. is taken by variable instance, which is represented by the form
- Here is important to explain the difference between the wa:SubFormAdder and its two subclasses, wa:ClassSelectorSubformAdder, and wa:InstanceSelectorSubformAdder.

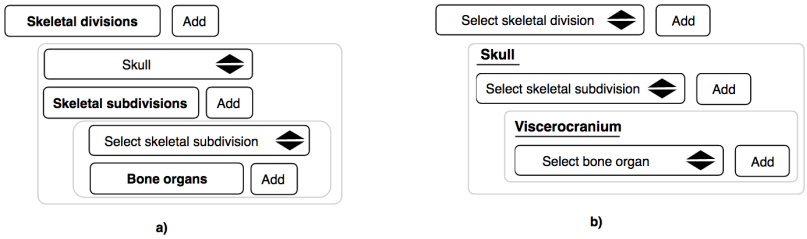

Figure 4.4: Difference between sub form adders

- min, max cardinality will be here addressed

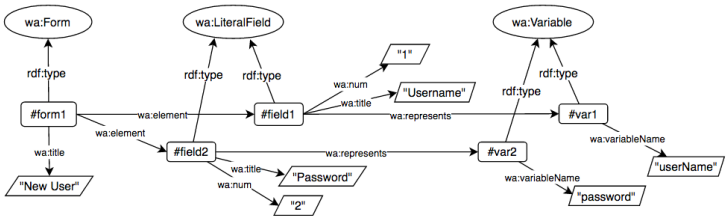

Figure 4.5: Form descriptor RDF Data

4.1.3 Data definition

- The goal of the data definition data scheme is to store what kind of triples has to be created
- The triples are denoted in the RDF vocabulary as statement
- There are two types of statements - the restriction and the data statement

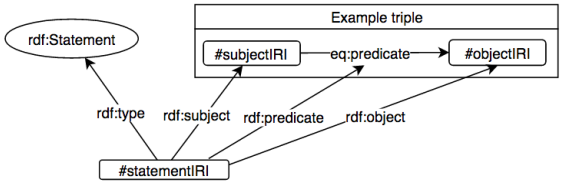

Figure 4.6: Statement in RDF Vocabulary

- And the nodes of the data graph to create is represented by the class wa:Variable
- These three classes represent the core of the ontology for data

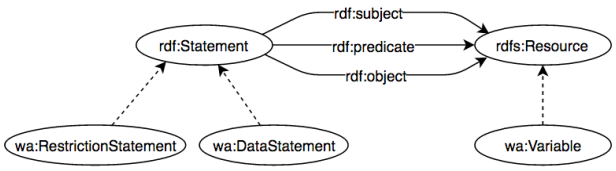

Figure 4.7: Core ontology

- The class wa:Variable is specified into Resource-, Class- and Literal-Variable
- And the LiteralVariable into further XML type variables
- Each variable have to have name, and they can hold a constant value or noted as main inputs if their value do not come from the form, but initially by the form loader HTTP request. In case of VIVO the main input variables are always the subjectUri, predicateUri, and rangeUri.

- There are two different types of statements, the class and the instance restriction statements.
- The class is used between classes, like which bone organ belongs to which skeletal division

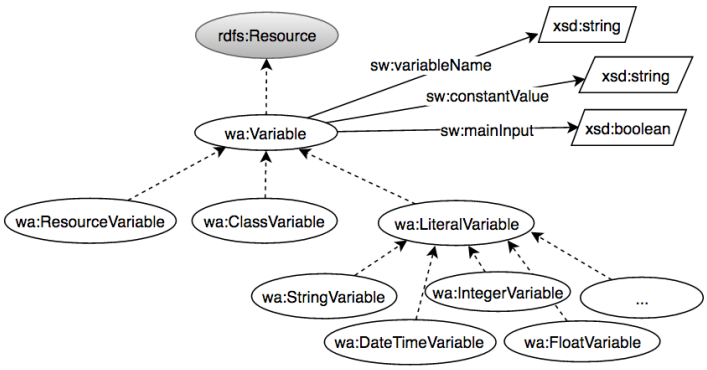

Figure 4.8: Variable types

- And the instance in turn expresses constraint between existing instances of the dataset to create, like the city-country relationship in the problem statement.
- The wa:TypeStatement is a statement, where the predicate is RDF type. It is interesting because it can act both as retriction and data statement as well. If the subject (so the instance) of the type statement appears on the interface, then the it is considered as restriction, but if the object (the class), then the it is data statement.
- Moreover the multi triple is really important because through them can be the multi dimensionality of the form expressed.

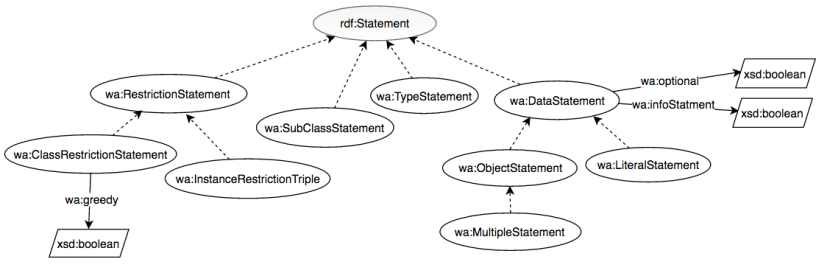

Figure 4.9: Statement types and attributes

- The following image show the configuration dataset for the problem in the previous chapter.

- In the middle there is a multi statement which expresses that more addresses of to a user can be assigned.

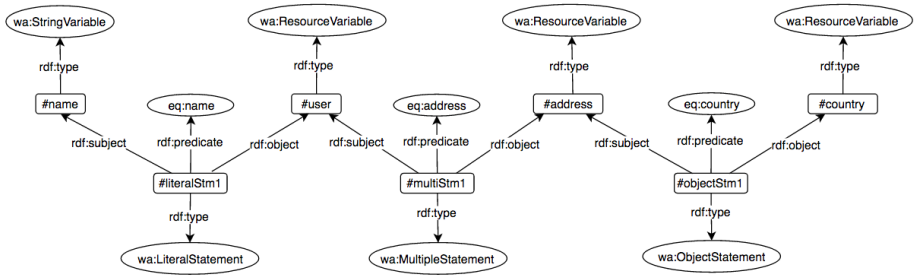

Figure 4.10: Statement configuration dataset I

- And here comes the type statement which is in this case acts as a restriction, and an instance restriction triple between city and country.

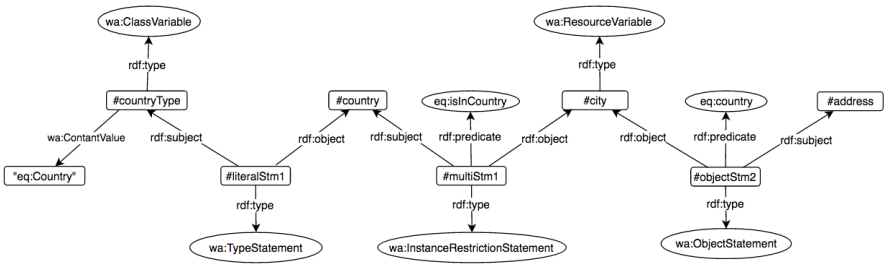

Figure 4.11: Statement configuration dataset II.

- The following image show the definition of the multi dimensional form.
- The literal field are not depicted for the sake of simplicity but they are connected to the #form1 and #subFormAdder instances

4.1.4 VIVO adoption

- As it was described in the previous section in VIVO each custom entry form is called by an HTTP request that contains three parameters: subjectUri, predicateUri and objectUri.

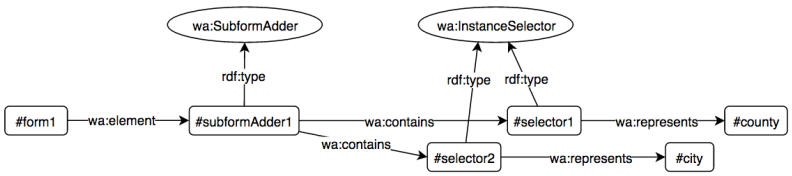

Figure 4.12: Form definition dataset

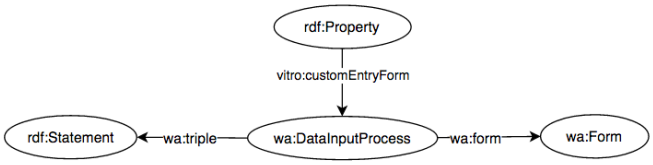

Figure 4.13: Base definition of the data input process

- So the key point is that the data input process instances can be queried by the property coming with the request.

4.2 Server-side implementation

- This section introduces how application on the server is able to operate based on data queried from the configuration dataset

4.2.1 Overview

- The following image shows the most important steps of the data input process

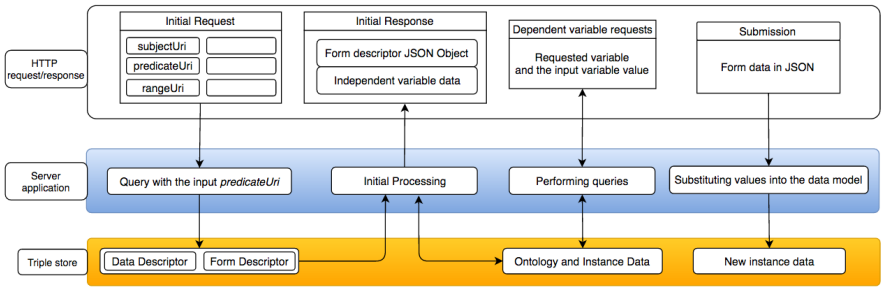

Figure 4.14: Overview of the mechanism on the server

4.2.2 Form representation in Java

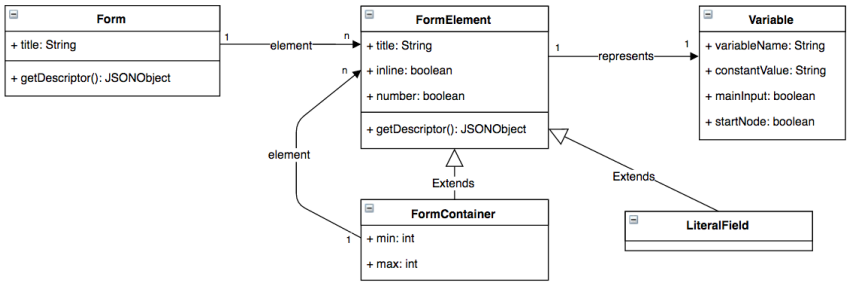

Figure 4.15: UML Diagram of the classes for the form

- The literal fields asks for the type of the variable it represents and the type of the descriptor will be based on this literal field.

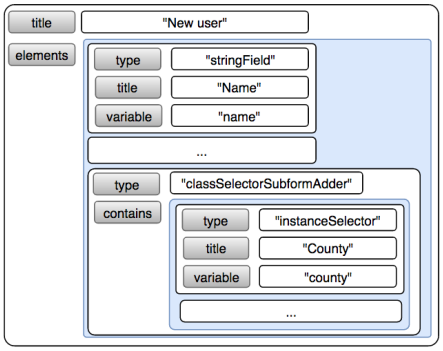

Figure 4.16: Form descriptor JSON object

4.2.3 Data model in Java

- Querying the configuration triples regarding statements
- Processing into the tree structured graph model

Validation:

- Data dependencies can be over graphs

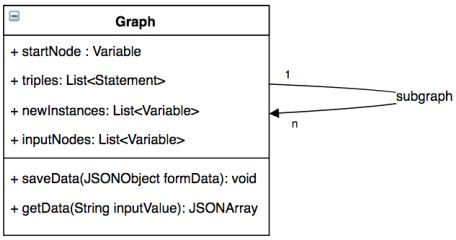

Figure 4.17: Graph UML

- But graphs can be connected only through multi statements
- The graph has to be a tree. There are no use cases right now where any loop would be required

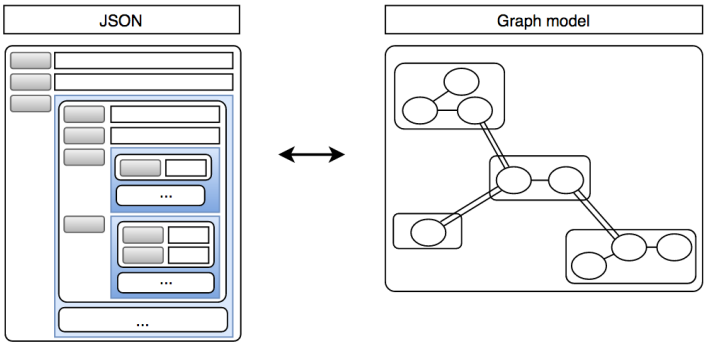

Figure 4.18: JSON vs graph model

- Data saving mechanism explanation
- Data retrieval mechanism explanation

4.2.4 Data Dependencies

- The data dependencies are important only for the form.
- Everything starts with the form descriptor.
- There are cases where from the main form no element appears on the form.

- A have to bring examples to some problems that illustrate the problem.
- Here comes at first the ontology awareness into question....

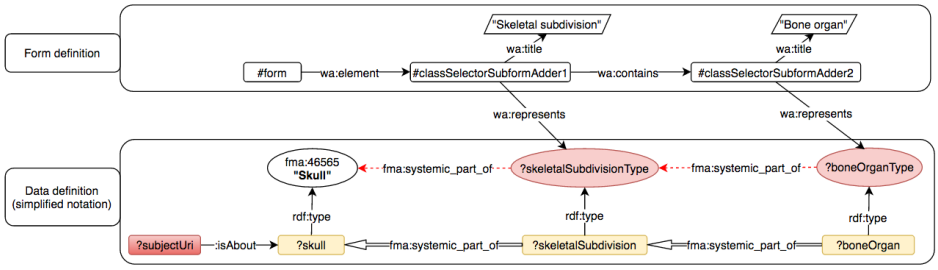

Figure 4.19: Example problem

where,

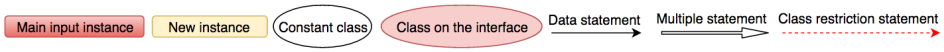

Figure 4.20: Elements of the simplified notation

Figure 4.21: Two restriction directions

- Post processing for different restriction types

4.2.5 Editing and deleting data

- This feature was not introduced on the overview image but it is really important.

4.2.6 Navigator

- Goal is to display more information about instances to select in order to facilitate their finding

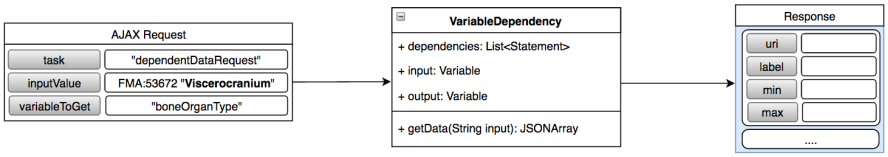

Figure 4.22: Getting dependent data

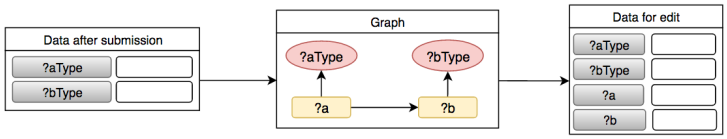

Figure 4.23: Difference between the submissions and edit data on the form

- By such advanced window it possible to offer filter options and loading the subset of the instances by introducing LIMIT on the SPARQL query. It could make the page more efficient
- But such element requires JavaScript code which could become complex
- Even if we need to offer some Navigator featured introduced in Section 3.2.3

- Example with RDF data configuration and routine for

4.3 Client-side implementation

- In this chapter it is discussed how the forms are implemented in JavaScript
- Each subsection contains code examples that gradually introduce the functionalities
- Codes are mostly simplified to facilitate the understanding

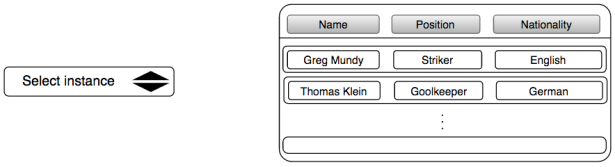

Figure 4.24: Selector VS form graph

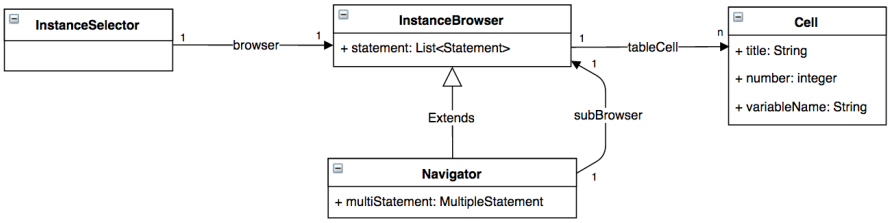

Figure 4.25: Navigator class diagram

4.3.1 Object oriented JavaScript

- The task of code in on the interface is to operate based on a configuration data dynamically.
- There two subtasks, the generation of the UI elements (i.e input fields, buttons, etc.) and manage the data input and display on the form
- To solve these problem, an object oriented approach is applied
- This means that there are classes that handles both the UI and the data related tasks
- See an example for the class definition in JS

```
class StringField {
  constructor(...){
    this.container = $("<div/>")
    ...
  }
  someMethod(){...}
}
```

Listing 4.1: JavaScript class

- Each form elements are represented by such objects, and the form loading based on the descriptor runs by the initialisation of these elements.

```
var formData = new Object()
for(var i = 0, i < formElements.length; i++){
  var descriptor = formElements[i]
  var element = null
  switch(descriptor.type){
    case "stringField":
      element = new StringField(descriptor, formData)
      break;
    case "..." :
  }
  $("#formContainer").append(element.container)
}
```

Listing 4.2: Form generation based on configuration data

- Explanation of the code ...
- This is the way how it is possible to generate interfaces

4.3.2 Handling data

- Above the element generation for the different forms it is necessary to handle of course the data based on the descriptor
- Each form element's descriptor contains a field called dataKey. The value of this field will be the key of data in the form data object.

```
class StringField {
  constructor(descriptor, formData){
    this.descriptor = descritor
    this.formData = formData
    this.inputField = $("<input/>").change(this.handler)
  }

  handler(){
     this.formData[this.descriptor.dataKey]=this.inputField.val
         ()
  }
}
```

Listing 4.3: Data saving

- The previous code illustrates how the form element object set the global form data field based on configuration data
- Further details about the code...
- Handling existing data
- It can be the case that form is loaded for editing. Then in this case the formData variable coming as input to the constructor contains the value for the dataKey?

```
class StringField {
  constructor(descriptor, formData){
    if(formData[descriptor.dataKey] != undefined){
      this.editMode = true
    }
  }

  handler(){
    if(this.editMode){
      var oldValue = this.formData[this.descriptor.dataKey]
    var newValue = this.inputField.val()
    }
    AJAX.updateField(oldValue, newValue)
  }
}
```

Listing 4.4: Data saving

4.3.3 Sub form adders

- Descriptor of the sub form adder contains a field called subform
- Then the initialisation of the form happend through the Form class.
- This is used by the initial form loading as well

```
class SubformAdder {
  constructor(descriptor, formData){
    ...
    this.addButton = $("<div/>").text("Add").click(this.add)
    this.subFormDescriptor = this.descriptor.subForm
    this.formData[this.descriptor.dataKey] = []
  }

  add(){
    var subformDataObject = new Object()
    this.formData[this.descriptor.dataKey].
      push(subFormDataObject)
    this.subFormContainer.append(
      new Form(this.subFormDescriptor, subFormDataObject))
  }
}
```

Listing 4.5: Sub form adder routine

- Important for each subform a new JSON object is generated which is pushed to the array
- The same way the subform adder checks if the this.formData[this.descriptor.dataKey] contains already existing data and adds them if they are there

4.3.4 Data dependency

4.3.5 Form validation

Appendix A

Glossary

Just comment `\input{AppendixA-Glossary.tex}` in Masterthesis.tex if you don't need it!

Symbols

$	US. dollars.

A

A	Meaning of A.

B

C

D

E

F

G

H

I

J

M

N

P

Q

R

S

T

U

V

W

X

Appendix B

Appendix

B.1 Something you need in the appendix

Just comment `\input{AppendixB.tex}` in Masterthesis.tex if you don't need it!

Erklaerung

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen/Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, bereits für eine andere Prüfung angefertigt wurde.

Ort, Datum	Unterschrift

Bibliography

[1] OWL 2 web ontology language document overview. W3C recommendation, W3C, October 2009. http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.

[2] Marcelo Arenas, Claudio Gutierrez, and Jorge Pérez. On the semantics of sparql. In *Semantic Web Information Management*, pages 281–307. 2009.

[3] Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, Michel Dumontier, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Alejandra Gonzalez-Beltran, Melissa A. Haendel, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Mark Jensen, Yu Lin, Allyson L. Lister, Phillip Lord, James Malone, Elisabetta Manduchi, Monnie McGee, Norman Morrison, James A. Overton, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Daniel Schober, Barry Smith, Larisa N. Soldatova, Christian J. Stoeckert, Jr., Chris F. Taylor, Carlo Torniai, Jessica A. Turner, Randi Vita, Patricia L. Whetzel, and Jie Zheng. The ontology for biomedical investigations. *PLOS ONE*, 11(4):1–19, 04 2016.

[4] Cornelius Rosse and José L.V. Mejino Jr. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6):478 – 500, 2003. Unified Medical Language System.