Using Semantic Web Technologies to Build a Community-driven Knowledge Curation Platform for the Skeletal Dysplasia Domain

Tudor Groza[1], Andreas Zankl[2,3], Yuan-Fang Li[4], and Jane Hunter[1]

[1] School of ITEE, The University of Queensland, Australia
tudor.groza@uq.edu.au, jane@itee.uq.edu.au
[2] Bone Dysplasia Research Group
UQ Centre for Clinical Research (UQCCR)
The University of Queensland, Australia
[3] Genetic Health Queensland,
Royal Brisbane and Women's Hospital, Herston, Australia
a.zankl@uq.edu.au
[4] Monash University, Melbourne, Australia
yuanfang.li@monash.edu

Abstract. In this paper we report on our on-going efforts in building SKELETOME – a community-driven knowledge curation platform for the skeletal dysplasia domain. SKELETOME introduces an ontology-driven knowledge engineering cycle that supports the continuous evolution of the domain knowledge. Newly submitted, undiagnosed patient cases undergo a collaborative diagnosis process that transforms them into well-structured case studies, classified, linked and discoverable based on their likely diagnosis(es). The paper presents the community requirements driving the design of the platform, the underlying implementation details and the results of a preliminary usability study. Because SKELETOME is built on Drupal 7, we discuss the limitations of some of its embedded Semantic Web components and describe a set of new modules, developed to handle these limitations (which will soon be released as open source to the community).

1 Introduction

Skeletal dysplasias are a heterogeneous group of genetic disorders affecting human skeletal development. Currently, there are over 440 recognized types, categorized into 40 groups. Patients with skeletal dysplasias have complex medical issues, such as short stature, degenerative joint disease, scoliosis or neurological complications. Since most skeletal dysplasias are very rare (<1:10,000 births), data on clinical presentation, natural history and best management is sparse. The lack of data makes existing patient cases a precious resource for biomedical research because they enable scientists to study, among other things, the effects of single genes on human bone and cartilage development and function. The resulting insights may lead to a better understanding of the pathogenesis of more common connective tissue disorders, such as arthritis or osteoporosis.

Unfortunately, due to the intrinsic complexity of dysplasias, correct diagnosis is often difficult. At the same time, only a few centres worldwide have the necessary expertise in diagnosis and management of these disorders. On the other hand, the identification of many skeletal dysplasia-causing genes and subsequent studies of their functions and interactions have led to an explosion in the knowledge of bone and cartilage biology. The biomedical literature now contains a large amount of information about individual genes and gene interactions, but it is often difficult to grasp how these interactions work together in a broader context (such as skeletal dysplasias). In turn, the focus on specific patient cases or genes makes it difficult to identify etiological relationships between skeletal dysplasias, or to recognise clinical or radiological characteristics that are indicative of defects within a specific molecular pathway.

The International Skeletal Dysplasia Society (ISDS)[5] has attempted to address some of these problems with its Nosology of Genetic Skeletal Disorders [1]. Since 1972, the ISDS Nosology lists all recognised skeletal dysplasias and tries to group them by common clinical-radiographic characteristics and/or molecular disease mechanisms. The ISDS Nosology is revised every 4 years by an expert committee and the updated version is published in a medical journal, being widely accepted as the "official" nomenclature for skeletal dysplasias within the biomedical community. While the content is invaluable, the format of the Nosology has several short-comings, including: (i) an inflexible classification scheme – each disorder being listed in one group based either on its clinical radiographic appearance or on its underlying molecular genetic mechanism; (ii) limited amount of cross-referenced information – each entry contains only the Online Mendelian Inheritance in Man (OMIM) number [2], the chromosome locus and the gene name, without being linked to widely used semantic data repositories, like the Gene Ontology [3] or UniProt [4], which would allow users to study further up-to-date relevant information; and most importantly, (iii) the lack of a shorter publishing cycle – the content quickly becomes outdated, as genes or disorders discovered after the publication date can no longer be included until the next revision (4 years later).

In addition to the above-mentioned Nosology issues, collaboration among experts is also adversely affected by a lack of an appropriate tool support. Currently, the community uses the ESDN (European Skeletal Dysplasia Network) Case manager[6] and Google mailing lists to share information and to exchange and discuss patient cases. Neither of these provides an ideal collaboration environment. While ESDN provides a structured (form-based) discussion forum to support the diagnosis process, mailing lists are merely long threads of free text. Leaving aside the complete lack of any formal representation or semantics, a major issue is the inability to transfer knowledge or provide links between the rich pool of patient reports and the ISDS Nosology.

In this paper we report on the efforts of the SKELETOME project[7], which aims to develop a community-driven knowledge curation platform for the skele-

[5] `http://www.isds.ch/`
[6] `https://cm.esdn.org/`
[7] `http://itee.uq.edu.au/~eresearch/projects/skeletome/index.html`

tal dysplasia domain. The SKELETOME platform[8] introduces an ontology-driven knowledge engineering cycle that supports the continuous evolution of the knowledge captured in the ISDS Nosology from existing patient studies, thus transforming into a living knowledge base. Concurrently, this knowledge informs the collaborative decision making process associated with newly arriving cases. Moreover, the underlying SKELETOME ontologies represent a foundational building block for linking to external resources and a mechanism for facilitating knowledge extraction and reasoning. SKELETOME is being developed by extending Drupal 7[9] with additional Semantic Web components to enable seamless and semantic-aware collaborative input, sharing and re-use of data and information among the experts in the field. The knowledge engineering cycle, together with the set of new Semantic Web Drupal modules (and some lessons learned from the existing ones) represent the main contributions of this paper.

The remainder of the paper is organized as follows. Section 2 describes the representational and functional requirements supporting the SKELETOME platform. Section 3 provides a detailed overview of the SKELETOME components and information flow. In Section 4 we discuss the preliminary evaluation, and before concluding in Section 6, we analyze some of the existing related efforts in Section 5.

2 Requirements

Since genetic disorders are typically quite rare, a global network of patients, clinicians and researchers is necessary to accumulate the critical mass of data and knowledge needed to address some of the greatest challenges in medical genetics, i.e., the development of evidence-based clinical management guidelines, the study of genotype-phenotype[10] correlations and the identification of disease modifier genes. Skeletal dysplasias are an ideal topic for a global medical collaboration network as the number of medical conditions is relatively small and well defined and there is an existing, tightly-knit and motivated community of clinicians and scientists willing to contribute, share and exchange case studies, data, diagnoses and clinical information.

Recognition of this opportunity, led to the establishment of the SKELETOME project – a collaboration between information scientists, Semantic Web researchers and clinical geneticists, led by the University of Queensland. In addition to a Web-based framework for enabling and encouraging the international skeletal dysplasia community (researchers, experts, clinicians) to contribute content, the most important requirements for the project (which emerged from direct discussions with the community) are the following:

Common terminology. The diagnosis and management of skeletal dysplasias depends on highly specialised domain knowledge across a number of disciplines (radiography, genetics, orthopaedics), which is not easily comprehensible

[8] `http://skeletome.metadata.net/skeletome`

[9] `http://drupal.org/drupal-7.0`

[10] *Genotype* refers to the genetic information of an individual, while *phenotype* describes the actual observed properties of an individual, such as morphology or development.

to individual communities or hospitals. In order to enable the exchange of knowledge between experts (across languages and disciplines), a common terminology is required, hence leading to a shared conceptualisation of the domain.

Data integration. Large datasets containing rich information on molecules (genes, proteins) already exist and the information relevant to skeletal dysplasias needs to be extracted and cross-referenced with the clinical data and knowledge produced by SKELETOME. The data cross-reference requires integration both at conceptual level, as well as, at actual data and instance level.

Privacy and access control. Actual patient studies and reports need to be visible only to the experts participating in the decision making process. Moreover, sensitive patient data (e.g., name, address, relatives) should only be accessible to the case initiator.

Knowledge transfer and sustainable knowledge evolution. The knowledge collectively acquired from the anonymized pool of patients represents a valuable asset from the conceptual perspective of the domain (materialized in the ISDS Nosology). Consequently, a seamless transfer of this knowledge is required to enable the dynamic and continuous evolution of the conceptual domain.

Capturing provenance and expertise. The contributed content may take several forms, ranging from personal observations to scientific publications. Independently of the form, SKELETOME requires a mechanism to keep track of the provenance of the data and knowledge, in order to ensure proper privacy and access control. It also needs to provide a measure of certainty of derived data and to leverage expertise from the content and to streamline the delivery of the most relevant information to the most appropriate person.

In order to support the above requirements, the SKELETOME platform provides the following services: (i) a collaboration environment for experts to exchange knowledge and patient cases and to build a repository of patient case studies linked to related evidence and Web resources (e.g., publications, radiographic data, gene databases, etc); (ii) a set of ontologies that capture the domain knowledge and underpin the platform; (iii) semantically enhanced content annotation and integration services; (iv) ontology-driven text processing of publications leading to rich semantic annotations; (v) enhanced image search and retrieval via ontology-based annotation; (vi) reasoning on anonymised patient data for semi-automated decision making.

3 The SKELETOME platform

The innovative aspect of the SKELETOME platform is the ontology-driven knowledge engineering cycle, introduced to bridge the current knowledge about the domain (partly captured in the ISDS Nosology) to the continuously growing pool of patient cases. The engineering cycle consists of two concurrent phases: (1) semantic annotation of patient instance data, and (2) ontology learning from patient instance data.

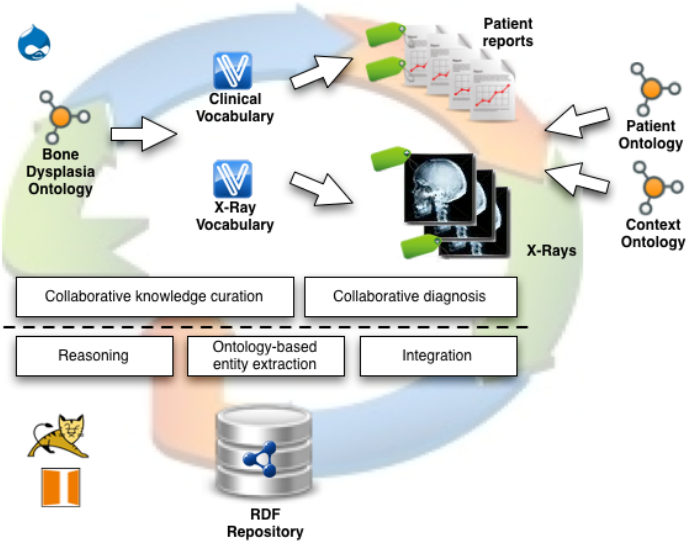

Fig. 1. The high level architecture of the SKELETOME platform.

We developed the Bone Dysplasia ontology[11] to overcome the short-comings of the ISDS Nosology and to describe the relations between bone dysplasias and the genotype and phenotype characteristics. The ontology is used to semantically-enrich patient reports and the associated X-Ray imagery. Additionally, in conjunction to two auxiliary ontologies (the Patient[12] and Context[13] ontologies), which capture patient and provenance information, we use the Bone Dysplasia ontology to enhance the collaborative diagnosis process. The resulting (RDF) instance data is then used in the reasoning process to propose novel genotype and phenotype characteristics to be associated to bone dysplasias, and hence support the collaborative knowledge curation and the evolution of the conceptual knowledge of the domain.

Fig. 1 depicts a high level overview on the SKELETOME architecture. The upper part of the architecture, including also the front-end, is developed using Drupal 7 and contains two main components (implemented via several Drupal modules): (i) the collaborative knowledge curation component, responsible for generating Drupal pages associated to ontology concepts, in addition to generating tagging vocabularies from the underlying ontologies, and (ii) the collaborative diagnosis component, responsible for capturing the information exchanged by the experts in the diagnosis process (i.e., diagnosis creation and rating or open discussions on diagnoses). The lower part of the architecture consists of the ontology-driven services, developed via a set of servlets hosted in Tomcat and using OpenRDF Sesame as RDF triple store. The Integration service bridges the Drupal world to the RDF back-end by managing several Drupal hooks on

[11] `http://purl.org/skeletome/bonedysplasia`
[12] `http://purl.org/skeletome/patient`
[13] `http://purl.org/skeletome/context`

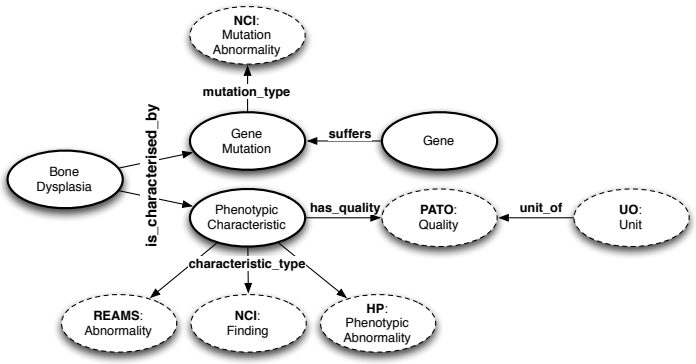

Fig. 2. A snippet of the Bone Dysplasia Ontology. The upper part of the ontology describes the genotypic information of bone dysplasias. The lower part relates bone dysplasias to phenotypic characteristics.

certain content types (or pages), e.g., Bone Dysplasia or Patient. Its role is to keep the RDF triple store in sync with the Drupal data, in addition to ensuring that no sensitive patient data is stored in the back-end. The other two services, i.e., Ontology-based entity extraction and Reasoning, have self-explanatory roles.

In the following sections we describe the underlying mechanisms used to develop the knowledge engineering cycle by means of the requirements introduced in the previous section. From a technical perspective we also identify some of the short-comings of the current Drupal 7 RDF support.

3.1 Common terminology and data integration

As mentioned in Section 1 the ISDS Nosology has a rigid structure and only partially covers the genotype information of the domain. More concretely, it merely lists the skeletal dysplasias, the genes responsible for the diseases and their locus, which leads to a poor description of the domain. Elements such as the gene mutation information and the radiographic or phenotypic characteristics are unfortunately ignored. For example, if we consider the *Stickler syndrome*, the ISDS Nosology only lists *COL2A1* as the responsible gene, and it does not mention that it might be caused by a *Missense mutation* in the gene (leading to a Glycine substitution with Arginine on position 219), or that some of the phenotypic characteristics are *Myopia* and *Cleft palate*, or that radiographically it can be characterized by *Dolichocephaly*.

To overcome these issues, and to extend the existing common terminology used by the community, we developed the Bone Dysplasia ontology (that defines more than 1200 concepts) to capture all the relevant knowledge by integrating and re-using well known ontologies, such as NCI Thesaurus [5], Human Phenotype Ontology (HP) [6] or the REAMS ontology[14] – describing radiographic features. Fig. 2 depicts a snippet of the ontology, showing the relation between

[14] `http://d-reams.org/?page_id=84`

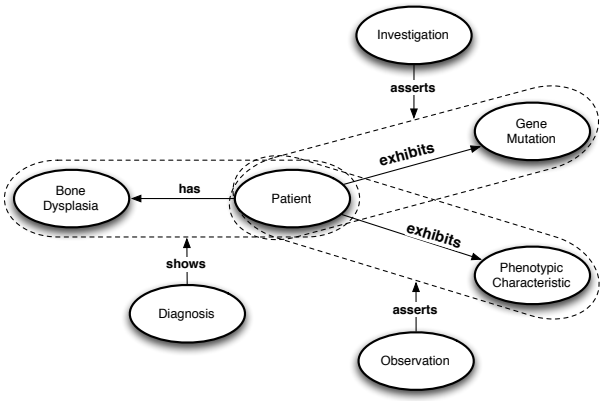

Fig. 3. A snippet of the Patient Ontology. The direct relations between Patient and the other three concepts are reified in order to capture the context in which their are materialized.

the root **Bone Dysplasia** class (further sub-classed by 40 bone dysplasia groups and then by specific skeletal dysplasias) and the **Gene Mutation** and **Phenotypic Characteristic** classes. As opposed to the ISDS Nosology, our **Gene Mutation** class provides the ground for encoding richer information about the characteristics of the mutation, e.g., type, position, original and mutated content. Similarly, the **Gene** class is linked (via annotation properties) to OMIM, MesH, UMLS and UniProt. These are only two examples where the ontology accommodates extended domain knowledge when compared to the ISDS Nosology.

In addition to the Bone Dysplasia ontology we have also developed a Patient ontology and a Context ontology. The Patient ontology captures knowledge about specific patient reports, hence describing "instances" of genotypic, phenotypic and radiographic characteristics of bone dysplasias in particular patients. As can be observed in Fig. 3 a **Patient** may *exhibit* diverse **Gene Mutations** or **Phenotypic Characteristics** which are *asserted* by **Investigations** or **Observations**. Similarly a **Diagnosis** *shows* that a **Patient** may *have* a particular **Bone Dysplasia**. The Context ontology is used to model the provenance of the patient information, including, for example, who suggested an **Investigation** or made an **Observation**, who and where a **Diagnosis** is documented, or even when a **Patient** exhibited certain **Phenotypic Characteristics**.

3.2 Knowledge transfer and sustainable knowledge evolution

The knowledge engineering cycle briefly introduced in the beginning of Section 3 was specifically designed to support this requirement. The first phase of the cycle uses the above-described ontologies to enrich the content created by the experts and to support the collective diagnosis process. This phase has, in reality, two sub-phases: (1) a sub-phase dealing with the evolution of the generic domain knowledge, i.e., classification of bone dysplasias and their descriptions,

and (2) a second sub-phase covering the actual use of ontologies for semantic annotation. The second sub-phase processes the patient instance data (i.e., semantically-annotated reports and diagnoses) to propose novel findings about bone dysplasias.

(1) Domain knowledge maintenance and evolution.

From a functional perspective the experts need to keep the domain knowledge up-to-date. SKELETOME publishes all the bone dysplasias and associated information (e.g., Genes) as Web pages (via specific Drupal content types). Hence each bone dysplasia has its own publicly available Web page, similar to the way in which Wiki systems work (see, for example, `http://skeletome.metadata.net/skeletome/bonedysplasia/achondroplasia`). The Bone Dysplasia ontology acts as a backbone for this set of pages, as they are automatically generated from (and are in sync with) the concepts defined in the ontology. The page generation is realized via a Drupal module that we have developed, in conjunction with the Integration service from the RDF backend. This module will be released as open source and may be useful to anyone who wants to build and maintain an ontology-driven content management site, starting from an existing ontology.

Currently, the Drupal RDF extensions allow one to map existing content types to ontological concepts and/or properties. Creating pages of those content types will result in Drupal creating the associated concept instances (via `rdf:type`). We were, unfortunately, unable to use this support for two reasons. Firstly, while the generic Bone Dysplasia content type could have been mapped to the corresponding ontological class, we required all its instance pages to represent classes themselves, and not instances (i.e., via `rdfs:subClassOf`). For example, the Web page about the **Stickler syndrome** should be mapped to the **Stickler syndrome** class in the ontology, and not to an instance of the **Bone Dysplasia** class. At the same time, mapping manually over 1000 concepts, currently present in the ontology, is neither feasible, not sustainable. Secondly, **Bone Dysplasias** are related to **Genes**, also modelled as content type instances. Instantiating node reference property values between custom content types is currently not supported by the RDF extensions. The need to provide reasoning support, which intrinsically requires explicit relations between instances, forced us to maintain all the RDF in the RDF back-end and develop specific Drupal hooks to keep the store constantly updated, via the Integration service.

In order to support maintenance and evolution of the knowledge base, SKELETOME provides support for adding, renaming or removing dysplasia groups (which are direct subclasses of the **Bone Dysplasia** class), moving bone dysplasias (which are direct subclasses of the groups) between groups, adding newly discovered gene mutations or phenotypic characteristics, hence manipulating the structure and content of the Bone Dysplasia ontology, without the experts being aware of it. Hiding the underlying ontological concepts and details was an easy decision, because the vast majority of the experts are simple computer users.

To ensure quality control over the content of the Bone Dysplasia pages, we have imposed an editorial process. Each bone dysplasia has an associated editor, responsible for keeping the explicit information and related knowledge up-to-date, by reviewing input from the community. In addition to scientific publi-

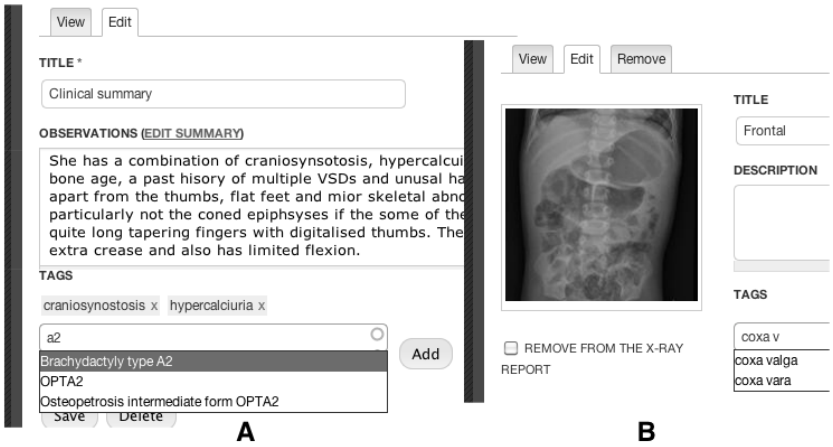

Fig. 4. [A] Semantic annotation of clinical summaries using a mixture of terms from Clinical Summary Vocabulary describing phenotypic characteristics; [B] X-Ray imagery tagging using terms from the X-Ray Vocabulary describing radiographic characteristics.

cations, this community input takes the form of statements asserted about the disease on its page, which can then be discussed and commented on by the entire community (thus enabling a "*wisdom of the crowds*" approach). On a periodic basis, the editor will incorporate in the main disease description those statements that were accepted by the community, and have reasonable support though scientific evidence.

In practice, the statements (or micro-contributions) carry a dual role: (i) they enable the transfer of knowledge from patient cases to the conceptual domain knowledge, as typically they would report on novel findings about the dysplasia from a particular pool of patients, and (ii) they allow us to create and maintain an expertise profile of the authors, which will lead to an authorship reputation system, similar to the one in WikiGenes [7]. The reputation of an author-expert is calculated based on the acceptance of her statements by the community, and hence the extent to which her contribution impacts on and advances the field.

(2) Semantic annotation of patient cases.

The Bone Dysplasia ontology provides not only the backbone for the evolution of the domain knowledge, but also the means for enriching patient cases with semantic annotations. Our goal is to provide experts with the mechanism for annotating both clinical summaries, as well as X-Ray imagery with domain concepts. In order to realize the annotation, we implemented a Drupal module that transforms a given ontology into a Drupal taxonomy (i.e., vocabulary) that enables tagging. The vocabulary import may be invoked from particular root concepts and can traverse the ontology up to a specified level or the leaf nodes. The actual tags are created by looking at the literal values of specified properties. For example, one may choose to create tags from *rdfs:label*, but also from *skos:altLabel*. One significant aspect of this module is that the generated tags retain a relation to the URI of the originating concept. Hence, when an expert

tags an X-Ray with a particular tag, s/he actually annotates the image with the ontological concept supporting the tag. This module will also be released as open source later this year.

Within the context of SKELETOME, this module is used to generate two vocabularies, as depicted in Fig. 1: a vocabulary for annotating clinical summaries (from the Bone Dysplasia ontology, HP, NCI and PATO, Phenotype and Trait Ontology [8], ontologies) and a vocabulary for annotating X-Ray imagery (from the HP and the REAMS ontologies). This distinction was specifically requested by the community in order to support their current terminological practice. The annotation of clinical summaries can be done manually or semi-automatically. The semi-automatic annotation is implemented by integrating the NCBO annotator [9] for entity extraction (via the Ontology-based entity extraction component of the backend).

The annotation of patient resources is transformed in the backend, via Drupal hooks, into relations between the patient instance and the corresponding concept instances. For example, consider the annotation depicted in Fig. 4: the patient instance being examined would be related via the *exhibits* relation to **HP:cranyosynostosis** and to **REAMS:coxa-valga**.

The collaborative diagnosis process works in a similar manner as the annotation of clinical summaries. By adding a diagnosis to a patient, in reality, the expert annotates the patient case with the corresponding Bone Dysplasia concept. In the backend, this translates into a reification of the relation between the patient instance and a particular Bone Dysplasia instance (see Fig. 3), which is then related to a freshly created **Diagnosis** instance that has context information attached to it. This context information is generated from the discussion among experts and includes the votes cast on the diagnosis.

Reasoning and searching on patient data.

Semantically annotated patient cases create a wealth of knowledge that represents a perfect application for reasoning. Leaving aside the access control aspects (detailed in the next part of the section), our goal is to close the knowledge engineering cycle by supporting both the evolution of the conceptual domain, as well as the collective decision making process in two ways. Firstly, we want to apply reasoning across current cases to propose diagnoses on newly published cases. Secondly, we want to infer novel findings on the conceptual domain, by focusing on the similarities between phenotypic, radiographic and genotypic characteristics in patient cases sharing the same diagnosis. With respect to this latter goal, we want to avoid discovering the obvious, e.g., that all patients diagnosed with *Achondroplasia* have a mutation in the *FGFR3* gene – this information is already in the domain knowledge.

The complexity of the two tasks requires a thorough investigation of the most suitable mechanisms to support them. Initially we considered reasoning across the instance data using SWRL rules for both tasks, however, we quickly realized that the rigidity of rule-based inferencing will not help us in fulfilling our goals. The SWRL rules had especially negative consequences on our second goal – the identification of features that are not present in the vast majority of cases (features that you would expect to be inferred via reasoning). Both diagnoses

and the presence of phenotypic characteristics cannot be stated with a 100% certainty. As a result, the collaborative voting mechanism currently featured in SKELETOME does not record a simple Yes/No, but uses a 5-star rating, hence allowing the experts to associate a level of uncertainty with their opinion. This rating cannot be converted into rigid/strict rules. Consequently, we are currently investigating ways of encoding the diagnosis information using fuzzy rules, in addition to using uncertainty and/or statistical reasoning techniques.

Besides reasoning (currently under investigation), the use of ontologies enables SKELETOME to provide semantic search functionality, dynamically related items and faceted browsing. This last aspect is particularly important for expert users as it allows them to quickly filter search results based on criteria, such as: patient ids, phenotypic or genotypic characteristics. Additionally, for an even richer browsing experience, we have integrated dynamic links to related knowledge items in some of the views (e.g., the dysplasia descriptions and patient clinical summaries). For example, a bone dysplasia description might have suggested links to patients diagnosed with this dysplasia or to related phenotypic characteristics. From a technical perspective, this is realized by following the relations in the ontology for the instance under scrutiny, and secondly, by analyzing the textual content (where possible) and extracting and linking domain concepts present in the knowledge base.

3.3 Privacy and access control

The information captured in SKELETOME is accessible via four layers of privacy and access control policies. The generic conceptual knowledge of the domain (i.e., the bone dysplasia pages and associated resources) are publicly available. The rest of the knowledge is private and accessible via group and individual-based access controls. Different groups of experts are registered within the platform and act as sub-communities within the greater community. The reason behind this group division is the need to share patient information only with a specific set of experts. Experts can, nevertheless, be members of multiple groups, and hence share their information and knowledge across all of them.

Sensitive patient information, such as name or address, is accessible only to the case initiator (individual-based access control). The purpose of exchanging patient cases is to foster advances in the field and to take advantage of the community-driven diagnosis process. However, sensitive patient information is not relevant for the diagnosis process, and hence is maintained only for provenance purposes. In reality, the so-called "participation" of the patients in these community exercises is acknowledged by the patients via written consents (also maintained within the platform, and included within the patient information).

As described in the previous sections, each patient's semantically annotated clinical data (including annotated X-Ray imagery) is stored and processed in the RDF backend. In order to enforce the individual-based access control over sensitive information, yet take advantage of the wealth of knowledge present in the entire pool of patient cases across all groups, we followed the principle of separation of concerns [10] (the fourth layer of access control). Drupal hooks were

implemented on the patient content type to filter the fields that are transformed into RDF instance data via the Patient and Context ontologies, and then stored via the Integration backend service. As a result, the RDF triples will model strictly phenotypic, radiographic or genotypic information, while the rest of the information (including the sensitive data) remains stored only in the Drupal database, and is subject to the three access restrictions described above. This allows us to perform reasoning on the entire set of patient clinical data, hence taking advantage of both the quantity and quality of the knowledge created by experts – whilst still restricting access to sensitive data.

3.4 Capturing provenance and expertise

The previous sub-sections have already provided an insight into the mechanisms implemented by SKELETOME to capture provenance and expertise. The Context ontology is used to capture provenance information, ranging from the author names and dates of assertions to timestamps on diagnoses or phenotypic characteristics. This information is then used to generate expertise profiles from asserted statements, forum discussions and collaborative diagnoses.

Currently, our expertise modeling is based on the mining of micro-contributions, based on a bag-of-concepts approach by aggregating concepts extracted via the NCBO annotator from any micro-contributions. Following this approach leads to several issues, especially when dealing with qualities of the phenotypic characteristics. For the future, our plan is to filter the output of the annotated entities by organizing them into tensors expressing quality – phenotypic characteristic associations and then use the tensors to compute the weights of the domain concepts in the context of both local and global contributions of particular individuals. Moreover, we also plan to take into account the processes that the micro-contributions undergo during their lifespan, e.g., how many times they were altered, the extent of alteration and whether or not they were incorporated in the main disease description.

4 Preliminary evaluation

We performed a preliminary usability study of the SKELETOME platform with a small group of eight experts from the community. The goal of the study was to compare the usability of SKELETOME against the two other "systems" currently used by the community, i.e., ESDN and Google mailing lists. At the same time we also wanted to understand how easy it is for the experts to adapt to using SKELETOME.

The evaluation consisted of two parts, with no training provided: (1) performing a series of operations on both SKELETOME and ESDN or Google mailing lists, and (2) completing a questionnaire about the usability of SKELETOME.

The tasks required to be performed for the first part were the following: (1) **Search & Browse:** search for a particular patient based on a given set of

Table 1. System usability questionnaire

Question
I found SKELETOME easy to use
I found SKELETOME to be unnecessarily complex
I think I require technical support to be able to use SKELETOME
I found the various features of SKELETOME to be well integrated
I think most colleagues would learn SKELETOME quickly
I felt very confident using SKELETOME
I needed to learn a lot about SKELETOME before I could effectively use it

phenotypic characteristics; (2) **Patient case manipulation:** upload and annotate a new patient case (i.e., clinical summary + 5 X-Ray reports); (3) **Collaborative diagnosis:** participate in the collaborative diagnosis process on a given patient; (4) **Domain knowledge manipulation:** modify the description of a bone dysplasia and add statements about it. This operation had to be performed only on SKELETOME (as the other systems do not have support for it). The experts were asked to pick one system in each category and motivate the choice by highlighting (via free text input) both the positive and the negative aspects.

The questionnaire required for the second part contained seven questions (see Table 1) with answers on a 5-point Likert scale ranging from "Strongly disagree" to "Strongly agree". These questions were adapted from the evaluation of the iCAT system [11] and from the System Usability Scale (SUS) [12].

The results of the evaluation were very positive. In the first part (choosing the "best" system), SKELETOME outperformed its opponents. In the **Search & Browse** task, the dynamically related items and the ontology-based faceted browsing and filtering of the search results were found to be very useful by all experts, and judged as critical missing aspects from the other systems. The **Patient case manipulation** task was the highlight of the evaluation, as all eight experts were particularly impressed by the possibility of annotating X-Rays and clinical summaries with domain concepts, and by the drag'n'drop functionality for uploading X-Rays. The third and last common task was found to be very similar to the ESDN functionality (50% voted for ESDN and 50% for SKELETOME), although the 5-star rating system was regarded as a positive addition by six out of eight experts. Finally, the fourth task (with no competition) was regarded as extremely useful to support their collaborative effort of maintaining and evolving the domain knowledge.

The questionnaire results were also positive. SKELETOME was found to be easy to use and well integrated by 87% of experts (7 out of 8), although 75% of experts did fell that they would require some time to learn to use it effectively. Functionalities, such as drag'n'drop or related items, made all the experts confident in using the platform. Some technical support was required with some users (37% of experts), however, it was only at the very beginning and did not influence the positive results of the questionnaire.

Overall, SKELETOME performed as we have expected, and indicated that Semantic Web technologies have the potential to make a real positive differ-

ence when applied in the right context and seamlessly built into familiar user interaction components.

5 Related Work

The SKELETOME platform is built on the work performed by the initiators and developers of the Drupal RDF extensions (described initially for Drupal 6 in [13]). Their continuous efforts, that resulted in support for RDF in core Drupal 7, are highly appreciated. However, as discussed earlier in the paper (see Sect. 3.2), in order to support a dynamically evolving skeletal dysplasia knowledge base, we required a different approach than the current top-down RDF mappings. Hence, we used the Drupal RDF core support, but developed our own set of modules, to be openly released to the community later this year.

The literature contains numerous descriptions of related systems, many using Wikis as their backbone. Semantic MediaWiki [14] was a pioneer in the area, being one of the first Wikis to embed semantic capabilities, packaged as extensions of MediaWiki. This led to its wide-spread adoption and application in many diverse domains. [15] is one example that adopted and applied the principles of Semantic MediaWiki. IkeWiki [16] on the other hand, is a stand-alone Semantic Wiki, providing similar functionality to Semantic MediaWiki and developed entirely in Java and AJAX.

Focusing on the biomedical domain, we identified BOWiki [17] and ConceptWiki [15] [18] as the most relevant with respect to our platform. BOWiki is a Semantic Wiki, designed for expert database curation, providing users with automated reasoning capabilities to verify the consistency of continuously added content to the knowledge base. ConceptWiki, or more concretely the WikiProteins part of it, is based on MediaWiki and enables experts to collaboratively curate knowledge about proteins. It incorporates several large knowledge bases, such as Gene Ontology, UMLS or Swissprot [19], to be used for the annotation and definition of terms, however, without providing a strict formalization of the knowledge or any reasoning support. The skeletal dysplasia curation aspect of the SKELETOME platform is reasonably similar to these approaches. However, while the generic goal of such Wikis is ontology engineering and population, SKELETOME extends this goal via the knowledge engineering cycle to learn new knowledge both from the growing pool of patient studies, as well as from the collaborative decision making process.

Another system of particular relevance is WikiGenes [7]. As opposed to typical Wikis, WikiGenes shifts the focus from creating knowledge to capturing the context of the knowledge via scientific artefacts (e.g., hypotheses or claims), in addition to fine-grained provenance. WikiGenes is special because it supports a reputation system for authors of the scientific artefacts based on their contributions to the field and their rating from other researchers. We adopted this concept and implemented it via the statements that can be added in conjunction

[15] `http://conceptwiki.org/`

to Bone Dysplasia concepts. However, in our case, these are more than mere conjectures, as they are (usually) supported by evidence emerging from the patient studies thus enabling one to track their evolution from the original patient data to the hypothesis at the generic conceptual level.

Outside of Wikis, the most notable and recent related effort is the custom WebProtégé [20] system built by the Stanford Center for Biomedical Informatics Research to help develop the 11th revision of the International Classification of Diseases (ICD-11) [11]. As with all the other tools in the Protégé suite, this system is specifically tailored towards efficient collaborative ontology engineering. As a result, in this respect it provides superior functionalities when compared to the Bone Dysplasia engineering aspect of SKELETOME. Nevertheless, we don't regard this as a negative point since it represents only one of the steps in the platform's knowledge engineering cycle. Additionally, in its current shape, SKELETOME serves its purpose of hiding the actual ontology evolution from the experts yet providing the mechanisms for keeping the knowledge up-to-date and enabling a shorter publishing cycle for the ISDS Nosology.

6 Conclusions

This paper describes the results to-date of our on-going effort in building a community-driven knowledge curation platform for the skeletal dysplasia domain. SKELETOME deploys an ontology-driven knowledge engineering cycle, aimed to support the evolution of the domain knowledge through the semantic enrichment of patient cases and via reasoning support that enables faster discovery of new knowledge and relationships. As the evaluation has shown, SKELETOME generates many benefits to the community and improves collaboration and knowledge exchange among the experts in the field. From a technical perspective, we believe that one of SKELETOME's main contributions is advancing the work started by Corlosquet et al. [13] in integrating Semantic Web technologies in the widely adopted Drupal CMS.

Future work on the platform will focus on developing novel mechanisms to support expertise modeling from micro-contributions and reasoning using fuzzy statements. From a functional perspective, we intend to integrate a notification mechanism, with personalized triggers on user-defined actions (e.g., notify me if anyone uploads a patient case that has these clinical attributes). In addition, to enhance the extent and ease of user interaction with the system, we plan to develop an email-based and iPhone app that enables the upload of clinical summaries and X-Ray reports.

Acknowledgments

The work presented in this paper is supported by the Australian Research Council (ARC) under the Linkage grant SKELETOME – LP100100156. The authors would like to thank Hasti Ziamatin and Razan Paul for their implementation support. Special thanks go to Tania Tudorache for her comprehensive and useful feed-back.

References

1. Warman, M.L., et al.: Nosology and Classification of Genetic Skeletal Disorders: 2010 revision. American Journal of Medical Genetics Part A **155**(5) (2011) 943–968
2. Hamosh, A., et al.: Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. Nucl. Acids Res. **33**(1) (2005) 514–517
3. Ashburner, M., et al.: Gene Ontology: Tool for the Unification of Biology. Nature Genetics **25**(1) (2000) 25–29
4. Bairoch, A., et al.: The Universal Protein Resource (UniProt). Nucleic Acids Research **33**(1) (2005) 154–159
5. Hartel, F.W., et al.: Modeling a description logic vocabulary for cancer research. Journal of Biomedical Informatics **38**(2) (2005) 114–129
6. Mabee, P.M., et al.: Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology and Evolution **22**(7) (2007) 345–350
7. Hoffmann, R.: A wiki for the life sciences where authorship matters. Nature Genetics **40** (2008) 1047–1051
8. Gkoutos, G.V., et al.: Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO. In: Proc. of the 31st Annual International Conference of the IEEE EMBS, Minneapolis, Minnesota, USA (2009) 7069–7072
9. Jonquet, C., et al.: The open biomedical annotator. In: Proc. of the 2010 AMIA Summit of Translational Bioinformatics, San Francisco, California, US (2010) 56–60
10. Dijkstra, E.W.: Selected Writings on Computing: A Personal Perspective. Springer-Verlag (1982)
11. Tudorache, T., et al.: Will Semantic Web Technologies Work for the Development of ICD-11? In: Proc. of ISWC 2010, Shanghai, China, Springer (2010)
12. Brooke, J.: SUS: a "quick and dirty" usability scale. In Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L., eds.: Usability Evaluation in Industry, London, Taylor and Francis (1996) 184–194
13. Corlosquet, S., et al.: Produce and Consume Linked Data with Drupal! In: Proc. of ISWC 2009, Chantilly, Virginia, US, Springer (2009)
14. Kroetzsch, M., et al.: Semantic Wikipedia. Journal of Web Semantics **5**(4) (2007) 251–261
15. He, S., et al.: Collaborative Authoring of Biomedical Terminologies Using A Semantic Wiki. In: Proc. of AMIA 2009 Symposium, San Francisco, California, US (2009) 234–238
16. Schaffert, S.: IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. In: Proc. of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, Manchester, UK (2006)
17. Hoehndorf, R., et al.: BOWiki: an ontology-based wiki for annotation of data and integration of knowledge in biology. BMC Bioinformatics **10**(S-5) (2009)
18. Giles, J.: Key biology databases go wiki. Nature **445** (2007) 691
19. Boeckmann, B., et al.: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**(1) (2003) 365–370
20. Tudorache, T., et al.: Supporting Collaborative Ontology Development in Protege. In: Proc. of ISWC 2008, Karlsruhe, Germany, Springer (2008) 17–32