The Potential for Use of Semantic Web Technologies in IK Management Systems

Thomas FOGWILL\textsuperscript{1}, Ronell ALBERTS\textsuperscript{2}, C. Maria KEET\textsuperscript{3}

\textit{CSIR Meraka Institute, Meiring Naude Rd, Pretoria, 0001, South Africa}
\textsuperscript{1}Tel: +27 12 841 3155, Fax: +27 12 841 4570, Email: tfogwill@csir.co.za
\textsuperscript{2}Tel: +27 12 841 3936, Fax: +27 12 841 4570, Email: ralberts@csir.co.za
\textsuperscript{3}University of KwaZulu-Natal, Private Bag X54001, Durban, 4000, South Africa
Tel: +27 31 260 1035, Fax: +27 31 260 7001, Email: keet@ukzn.ac.za

Abstract: This paper describes several opportunities to support and enhance Indigenous Knowledge digital repositories through the application of ontologies and semantic web technologies. Several needs for Indigenous Knowledge management systems are articulated. They are analysed to evaluate the possible application of semantic web technologies in IK management systems. Based on their potential for impact and their maturity, a number of these possible applications are recommended for further investigation and inclusion into current or new IK management systems. These include: an enhanced, ontology based querying facility; a facility to browse the knowledge using ontology navigation; a natural language independent ontology for multilingual data access; and a facility to support collaborative knowledge generation. The implementation of selected semantic web technologies in the National IK Management System (NIKMAS) is briefly described.

Keywords: Indigenous Knowledge Management Systems, Ontologies, Semantic Web Technologies

1. Introduction

Indigenous Knowledge (IK) is the local, traditional knowledge held by people of a particular area. It is central to their cultural heritage and holds significant value. The collection and management of IK is increasingly important for the purposes of preservation, protection, conservation and promotion [1]. IK in rural areas is mostly available in oral format and is held by IK holders in communities. Therefore, in this context, IK is mostly collected in free-text format which makes the management and useful dissemination of IK challenging.

An ontology is a logic-based, formal, precise description of a specific domain [2][3]. The domain is described in such a way that the meaning is understandable by humans and can be processed by computers. Without ontologies, computer systems treat new information as unknown content and retrieval and manipulation is performed through techniques such as pattern matching. With the formalisation of the domain through an ontology, the subject domain semantics (meaning) of the content becomes understandable to the computer system. This opens up many new possibilities for information storage, access, manipulation and processing using ontology-enhanced software systems. The application of automated reasoning services over ontologies allows the system to verify the logical theory’s consistency, derive knowledge, and intelligently query the data, information, and knowledge.
Semantic Web technologies are being developed to realise the aim of the Semantic Web by providing precise answers to complex queries over a network of information sources [4]. They are aimed at a number of goals, including [5]: creating a web of linked data, linking all information; enriching data with additional meaning to make it usable to people and machines; querying over a web of data sources; enabling reasoning over data through rules; and improving collaboration, research and development and innovation.

While the Semantic Web is still a work in progress, innovative tools and technologies have been developed that can be utilised in software development today. In particular, there are opportunities for utilising ontologies and semantic web technologies in the area of IK management systems in order to address some of the challenges in management and dissemination of IK.

The aim of this paper is to explore existing semantic web technologies, their maturity and the opportunities they create when applied in the domain of IK management. It will also describe the initial application of selected semantic web technologies in the IK management domain. The result of this study will be of interest to organisations involved in information systems for IK management and dissemination.

The paper is structured as follows: The needs identified for IK management systems that are potentially applicable to Semantic web technologies are described in Section 2. A survey of applicable semantic web technologies is presented in Section 3. The opportunities for application in the domain of IK management systems is described in Section 4. Section 5 contains a description of the application of selected semantic web technologies in the IK management domain. The paper is concluded in Section 6.

2. IK Management System Needs

The needs for IK management systems were investigated through workshops conducted with participants from government departments, higher education institutions, science councils, scientists, traditional authorities and community based organisations. The output of these workshops were analysed and filtered to yield the following list of needs that could potentially be addressed by semantic web technologies:

1. **Effective interrogation of IK**: The unstructured nature of IK and the fact that it often contains vernacular concepts makes interrogation and dissemination extremely challenging. In particular the ability to query IK, browse IK based on inherent structures and relationships and find answers to complex questions were identified as specific needs.

2. **Access to multi-lingual information**: IK is often collected in the native language of the IK holder. Automatic translation services are not yet readily available for all African languages and manual translation is expensive and time consuming. The ability of accessing and querying over information in different languages will be very beneficial in the context of IK management.

3. **Collaborative knowledge generation**: IK is a shared resource over individuals in traditional communities, geographical areas or communities of practice. A facility to capture and generate the knowledge in a collaborative fashion will increase the effectiveness of IK collection and management initiatives.

4. **Information classification**: Due to the unstructured nature of IK it is difficult to classify the collected knowledge correctly. Classification is required to effectively structure IK repositories.

5. **Formalisation of information**: IK is a valuable resource for responsible scientific discovery. The formalisation of applicable IK will enhance the effectiveness of scientific exploration.
3. Semantic Web Technologies

This section contains a description of existing semantic web technologies that may be applicable in IK management systems. The selection of semantic web technologies investigated were based on the needs described in Section 2.

3.1 Semantically Enhanced Querying

Semantic web technologies can enable enhanced querying of IK, by going beyond the simple string matching used in keyword-based search and using the semantics of the metadata stored with the IK. Keyword searches use string and linguistic matching of the search phrase and the information to be searched, but cannot exploit subject domain semantics and knowledge about the structure of information to find better results. Using ontologies allows the use of semantics in search, resulting in more precise and relevant results, even across institutional/software system boundaries.

Example: Plant A is used to treat “lung conditions” and plant B to treat “shortness of breath”. Searching for “treatments for lung conditions” in an ontology-driven system uses the relationship between “breathing” and “lungs” to return both A and B, while a standard keyword query will return only A.

Using reasoning services over ontologies also allows the system to compensate for missing (incomplete) information during the execution of a query. Thus, even if not all the information was explicitly captured in the system, it can, through inference, deduce the correct answer to a query. This is not possible in standard queries.

Example: Tuber T is a part of a Plant P. It is captured in the system that Plant P occurs in a particular geographical area, but it is not captured explicitly where tuber T occurs. A query can, through reasoning, find the location of tuber T, even though its location was not explicitly stated, by inferring knowledge from parthood (of the plant), and mereotopological and spatial theories (to deduce the location).

Examples of current research and technology include: Ontology-Based Data Access enabling access to data though a formal conceptual data model [6] and representing mereotopological relations in OWL [7].

Proofs of concept of this approach include: WONDER: a graphical tool to browse and query databases using an ontology [8]; Quelo: an intelligent query interface based on ontology navigation using pseudo-natural language [9]; GoPubMed: a tool to explore biomedical literature using the Gene Ontology [10]; Textpresso: an ontology-based information retrieval and extraction system for biological literature [11]; and DLMedia: an ontology-mediated multimedia information retrieval system [12].

3.2 Semantic Browsing of Information

An ontology is a formal representation of a subject domain that is independent from how the information is stored in the system and closely resembles the knowledge from domain expert's point of view. This creates the opportunity to have a facility to browse the information in the system based on the concepts defined in a domain and the relationships between them. Ontology-guided navigation will enable users to discover information based on the meaning of the topic and its relationship with other topics in the domain.

Examples of current research and technology include: MADS: a spatio-temporal model that provides a rich set of constructs for modelling data structures, spatial features, temporal features, and multi-representation features including an associated data manipulation and query language [13]; C-OWL: an approach to contextualise ontologies [14]; and SKOS: specifications and standards to support the use of knowledge organisation systems such as
thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web [15].

Examples of prototype implementations include: OntoGraf: a tool to interactively navigate the relationships of an ontology [16]; and Cytoscape: an open source bioinformatics software platform for visualizing molecular interaction networks [17].

3.3 Question Answering Facility

Question answering can provide specific answers to complex questions related to the knowledge in a domain and the underlying information in the system. This differs from the advanced searching described in Section 3.1, in that the question - posed in natural language - is also analysed to determine the type of question and an intelligent answer is compiled based on the knowledge in the ontology and the underlying data. Typically, the kinds of queries are different, as are the data sources, and the answers are approximations and thus have only a degree of accuracy.

**Example:** A question like “what part of plant X is mostly used for ailment Y?” can be answered through this type of facility with some degree of accuracy; the system will know that for a “What?” query, the answer must be a non-human object.

Examples of current research and technology include: Ontology resources in question answering [18]; Linguistic techniques for question answering including stemming and query expansion [19][20]; and Machine learning techniques [21][22].

Examples of proofs of concept applications include: Aqualog: a portable question-answering system which takes queries expressed in natural language and an ontology as input and returns answers drawn from the available semantic mark-up [23]; and Question answering in the Agricultural [24] and software testing [25] domains.

3.4 Multi-Lingual Access to Information

Ontologies are logical representations of a domain and can thus be natural language independent. This creates the opportunity to enable browsing and searching of knowledge in different languages and of accessing related information stored in different languages.

**Example:** A query can be formulated in any supported language. Internally, the query is translated into a query over the ontology, which is language-independent, and results can be extracted and presented in any of the available languages.

Examples of current research and technology include: query expansion for queries in different languages [26]; multi-lingual ontologies [27][28][29]; lexicalised ontologies [30][31]; annotation of information in different languages [32][33]; and parsers, morphological analysers and grammar engines.

Proofs of concept include: MUSIL: a multilingual search facility in a library [34]; Dogma: ontology learning supporting multiple languages [35][36]; the Monnet project: a project towards providing access to information across multilingual barriers by using a combination of Machine Translation and Semantic Web Technology [37]; and OntoVerbal-M: an ontology verbaliser that transforms OWL into fluent natural language paragraphs in multiple languages [38].

3.5 Collaborative Knowledge Generation

Semantic web technologies can be utilised to provide a facility to enable a community to create a precise representation of the knowledge in their area of interest in a collaborative manner. Content loaded can be tagged by the community to enrich the meaning and accessibility of the items and inform the definition of their metadata for that area of interest.
Example: A community of drum builders can collaborate to define the domain of their interest by collaboratively developing an ontology or conceptual data model of the domain. As content is added to the system, community members can use this shared representation to tag and annotate the content.

Current research and technology include: Modelling tools such as mind maps, graphs, conceptual models; and Folksonomies: enabling communities to classify digital assets through shared metadata [39].

Proofs of concepts and applications include: NeOn Toolkit: an ontology engineering environment with support for the ontology engineering life-cycle [40]; Freebase (http://www.freebase.com): a graph-shaped database of structured general human knowledge, inspired by Semantic Web research and collaborative data communities [41]; and Semantic Wikipedia: an enhanced Wikipedia with semantic technology to enhance the machine interpretability of the information and the links between information [42]. Two different implementations of the Semantic Wikipedia have been developed: Moki: a modelling environment enabling different actors to model an enterprise through collaboration [43]; and OntoWiki: a distributed knowledge engineering tool including a visual representation of a knowledge base, with different views on the related data [44].

3.6 Classification of Information

In order to utilise the full power of ontologies, new information received by the system must be accurately classified according to the defined concepts in the knowledge base. The accurate classification of information will enhance the comprehensibility of the knowledge and the accessibility and ease of retrieval of the information.

Examples of current research and technology include: Taxonomic classification: the classification of knowledge on a class level based on the declared properties in the knowledge base; Folksonomies: enabling communities to classify digital assets through shared metadata [39]; and Formal concept analysis: using a collection of objects and their properties to automatically derive an ontology or extend an existing knowledge base using the knowledge base itself together with information provided by a domain experts [45].

Examples of proofs of concepts and applications available include the automatic classification of bioinformatics instances based on declared properties [46].

3.7 Formalisation of Scientific Knowledge and Discovery

Formalising the knowledge and information captured in the system can enable scientific verification and discovery. The IK captured in the system can be formalised into an ontology and enriched with scientific information. This will allow scientists to use the knowledge for purposes such as hypothesis testing and consistency testing of theories.

Examples of proofs of concepts and applications available include: an automatic system for addressing the Chemical Compound problem, by interpreting transformations on the compound structures as updates in an ontology [47]; automated reasoning services for bioinformatics [48]; hypothesis testing using rough ontologies [49]; testing the differences of versions of a knowledge base using semantic diff [50]; and automatic classification of protein phosphates using an ontology, resulting in classification that surpassed that of human annotators and identified gaps in the theory that would not have been possible otherwise [46].
3.8 Data Integration

An aim of the semantic web is to integrate the information from different data sources through shared ontologies. This will enable users to search information over multiple information sources without being aware of where the information is physically located.

Examples of current research and technology include: instance-based data integration: integration of data at the level of the information rather than the knowledge base; and schema-level data integration; integration of data at the level of a shared ontology [51].

Examples of proofs of concepts and applications available include: data integration in the biomedical domain through shared ontologies [52][53]; and ontology-based semantic integration of biological databases [54].

4. Recommendations

The maturity and application readiness of the semantic web technologies described in the previous section are reflected in Table 1.

<table>
<thead>
<tr>
<th>Opportunity</th>
<th>Basic research results</th>
<th>Applied research in progress</th>
<th>Prototype / proof of concept</th>
<th>Applications in progress</th>
<th>Applications available</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Enhanced querying</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>2. Semantic browsing</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>3. Question answering</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>4. Multilingual information access</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>5. Knowledge generation</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>6. Classification of information</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>7. Scientific formalisation &amp; discovery</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>8. Knowledge-based data integration</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
</tr>
</tbody>
</table>

The maturity of the selected semantic web technologies can be summarised as follows:

1. Enhanced querying: Advanced, ontology-based search is an emerging field. There are numerous proofs of concept and prototypes but few robust and scalable solutions.
2. Semantic browsing: Tools that navigate the knowledge described by an ontology are mature; however, further work is needed to enable robust and scalable linkages to underlying data sources.
3. Question answering: Promising results have been achieved using linguistic and machine learning techniques in recent years. However, ontologies and semantic web technologies are still immature and need additional research in order to move towards mature tools and applications in the field.
4. Multilingual information access: The technology in this field is close to mature and has been successfully implemented in a number of applications.
5. Knowledge generation: The technology for knowledge generation is mature and a number of robust applications have been successfully developed.
6. Classification of information: Some experimental implementations have been developed in the bioinformatics domain with significant impact. However, technologies for broader domains must still be researched and developed.
7. Scientific formulation and discovery: The related technology is still low in maturity while the impact of success will be high.
8. Knowledge-based data integration: Technology for data integration has been successful up to the level of schema integration (e.g. IBM Federator); however, generic solutions based on Semantic Web technologies - at the knowledge level - are still immature and not scalable.

Based on an analysis of the maturity of the technologies and the diverse needs of IK management systems, it is recommended that the following ontology-based applications be
considered for implementation in IK management systems: (1) an enhanced, ontology based querying facility to significantly improve the standard keyword search on the information in the IK repository; (2) a facility to browse the knowledge captured in the IK domain ontologies and to explore the related information stored in the IK repository; (3) a language independent ontology to support multilingual data access; and (4) a mechanism to support collaborative knowledge generation in the IK management system.

5. Implementation

The National Recordal System, undertaken by the South African Government’s Department of Science and Technology, strives to capture, preserve and protect South African IK, using a National IK Management System (NIKMAS). The National IK Management System (NIKMAS) is an ICT tool supporting the NRS. NIKMAS represents a digital IK repository, where IK is described using multimedia and metadata. This repository is enhanced using selected Semantic Web technologies to support improved IK management and retrieval [55].

In particular, the information in NIKMAS is described and structured through an underlying domain ontology. The application of the ontology creates the opportunity to apply the technologies described in Sections 3 and 4. A prototype for an intelligent querying facility was developed, using a custom ontology developed for the IK themes of African Traditional Medicine and Food Security. This ontology was used to query annotated IK collected in the NIKMAS repository. Preliminary evaluation of the intelligent querying facility shows that the application of semantic web technologies increases the comprehensibility of the information in the repository and the effectiveness of querying over the IK in the repository.

6. Conclusions

Several needs for IK management systems were described. These needs were analysed to identify where the application of semantic web technologies could be relevant. The investigation identified the following possible opportunities for the application of ontologies and semantic web technologies in IK management systems: (1) Semantically enhanced querying; (2) Semantic browsing of information; (3) Question answering; (4) Multilingual information access; (5) Collaborative knowledge generation; (6) Classification of information; (7) Theme based exhibitions; (8) Formalisation of scientific knowledge and scientific discovery; and (9) Knowledge-based data integration.

Based on the potential impact and the maturity of the technology, the following opportunities are recommended for further investigation and inclusion into current or new the IK management systems: (1) An enhanced, ontology based querying facility; (2) A facility to browse the knowledge in NIKMAS using ontology navigation; (3) A language independent ontology for multilingual data access; and (4) A facility to support collaborative knowledge generation.

The implementation of an ontology and the development of an intelligent querying facility prototype in the National IK Management System (NIKMAS) was briefly described.

Further research will entail evaluation and extension of the intelligent querying facility prototype and implementation of more of the recommended Semantic Web technologies in the NIKMAS.

Acknowledgements

This work was made possible with the support of the South African Department of Science and Technology's National IK Systems Office and also by the support of the European
Union’s 7th Framework Programme Net2 project, which funded a collaboration visit to South Africa for C.M. Keet.

References


