2008-01-01

A Study on the Use of Ontologies to Represent Collective Knowledge

John McAuley
Dublin Institute of Technology, john@dmc.dit.ie

Recommended Citation

McAuley, J. (2008). *A study on the use of Ontologies to Represent Collective Knowledge.* Masters dissertation. Dublin Institute of Technology. doi:10.21427/D76P67 each question by comparing, connecting and mapping instances from the underlying domain ontology. Explorer's domain was, in contrast, represented along three axes, i.e. time, objects and monuments. The Explorer forum therefore provided users with the ability to query for narrative content within those three axes.

The Explorer approach to 3D visualisation

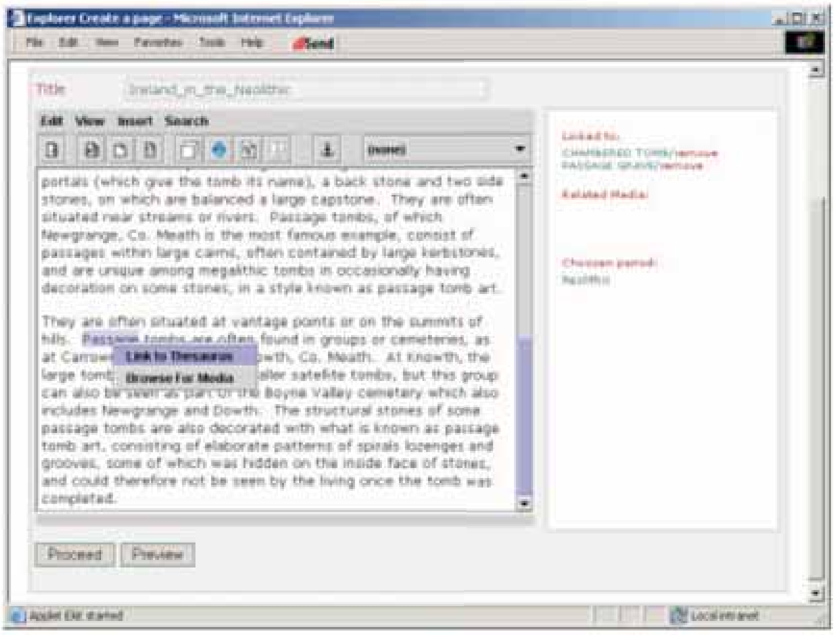
To accompany written narrative, a second tool was constructed, supporting users in creating 3D narrative presentations. Users could annotate VRML[85] objects with instances from the knowledge base or concepts from the domain ontology, and upload the VRML as a 3D presentation to the forum (Redfern and Kilfeather 2004). Forum members, when browsing the knowledge-base, could view and interact with the 3D presentation. The annotated VRML objects were clickable providing the user with an explanation of the object through narrative from the underlying knowledge base.

[85] Virtual Reality Modelling Language is a standard for representing 3-dimensional interactive vector graphics suited to use on the World Wide Web.

farming" and approaches the Explorer forum with the concepts "sickle"," cow", and "Dun Aonghasa" hill fort, in mind. The author writes the narrative text and relates the story concepts to terms from the underlying vocabulary. In this case, they relate the story to the term sickle in the adapted English Heritage objects thesaurus (see section 5.2), to the term hill fort in the adapted English Heritage monuments thesaurus (also section 5.2) and finally the time period Stone Age to the glossary of time periods as discussed in section 6.2. The author publishes the completed narrative presentation, making the narrative available to the community. The corollary of this procedural approach to narrative is text buttressed with terms from the underlying thesauri and glossary of time periods.

Both the approaches as implemented in the Bletchley Park and Explorer forums aimed to highlight the conceptual structure which underpins narrative. Approaching narrative in this way suggests that stories can be retrieved their conceptual structure rather than through simple keyword matching (as with keyword search). The narrative encompassing the three concepts in Figure 25, for example, may be retrieved by a user interested in Stone Age farming and who searches for any of the three concepts found in the underlying narrative base. Therefore, a user interested in Stone Age farming will, from the above example, retrieve stories based on any of the concepts, *Stone Age*, *Sickle* or *Hill Fort*. The merits of approaching narrative in this way are evident when compared to keyword search. This is because stories are retrieved based on the concepts they embody not the words they contain.

Within this context, the story fountain tool proposed a question driven approach to accessing the narrative base. The user starts with an open ended question and breaks that question into more discrete questions, each relating to specific domain concepts. Therefore the question driven interface supports the user in narrowing the results from vocabularies. The search results are displayed in a pop-up window. The interface was developed as a Java Applet with a set of server side components.



Title Ireland_in_the_Neolithic

portals (which give the tomb its name), a back stone and two side stones, on which are balanced a large capstone. They are often situated near streams or rivers. Passage tombs, of which Newgrange, Co. Meath is the most famous example, consist of passages within large cairns, often contained by large kerbstones, and are unique among megalithic tombs in occasionally having decoration on some stones, in a style known as passage tomb art.

They are often situated at vantage points or on the summits of hills. Passage tombs are often found in groups or cemeteries, as at Carrow[illegible]owth, Co. Meath. At Knowth, the large tomb[illegible]aller satellite tombs, but this group can also be seen as part of the Boyne Valley cemetery which also includes Newgrange and Dowth. The structural stones of some passage tombs are also decorated with what is known as passage tomb art, consisting of elaborate patterns of spirals lozenges and grooves, some of which was hidden on the inside face of stones, and could therefore not be seen by the living once the tomb was completed.

Linked to:
CHAMBERED TOMB/remove
PASSAGE GRAVE/remove

Related Media:

Choosen period:
Neolithic

Figure 24: The Explorer narrative authoring and annotation interface

Semantic browse and search

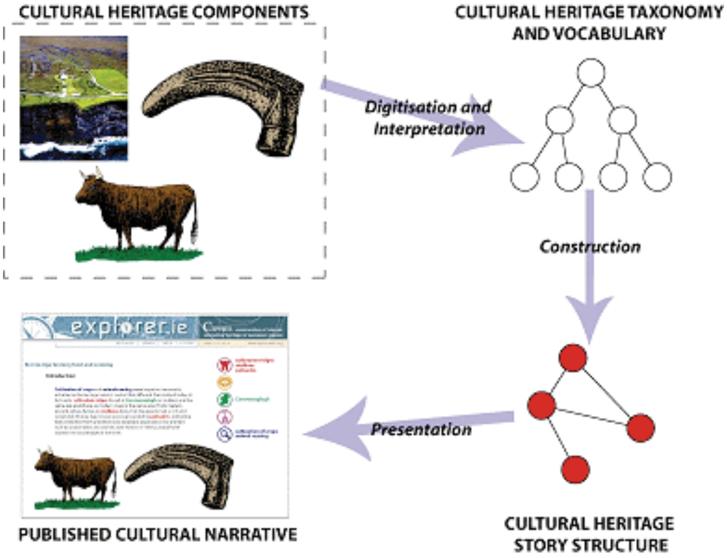

CULTURAL HERITAGE COMPONENTS
CULTURAL HERITAGE TAXONOMY AND VOCABULARY
Digitisation and Interpretation
Construction
Presentation
explorer.ie
PUBLISHED CULTURAL NARRATIVE
CULTURAL HERITAGE STORY STRUCTURE

Figure 25: An example of publishing a narrative in the Explorer forum.

Figure 25 illustrates an author constructing a narrative presentation for the Explorer forum. The author, in this case, wishes to write a narrative on “Stone Age

- **Interpretation of heritage resources:** The forum supports the interpreting phase of the knowledge cycle through the active annotation of new resources (image, film, sound). The annotation process involves a novel approach to the interpretation of resources with the domain ontology as presented in the following section.
- **Constructing of new stories:** The Story Creator supports the user construct new stories; semantic browse and search facilities help users retrieve concepts from the knowledge base, which they subsequently use to buttress their narrative content.
- **Juxtaposition of narrative:** *Story Trail Creator* was developed to allow forum members publish a linear narrative format through a collection of stories under a common theme.

Semi-automatic narrative annotation

The authoring interface (Figure 24) provided a means for users of the forum to intuitively create narrative content and annotate this content with terms from the underlying vocabularies. The author had the option of either allowing the system to propose annotation terms, an approach described as semi-automatic annotation, or manually choosing annotation terms. The first method analysed the complete narrative text, compared that text with the underlying vocabularies and then proposed terms from that matching process. The second approach allowed users to highlight text in the editor panel and using a right mouse click they could search for the word or phrase in the available vocabularies. The highlighted term is first put through a Porter Stemmer algorithm to remove stop words and improve the chances of a match in the

Other domain-specific differences occurred where terms in the EH and Irish thesauri were very similar in spelling but semantically dissimilar, such as the term Sheila-na-gig "A small carved figure, usually female in appearance, probably representing fertility charms, found on Romanesque churches in the West of England". This is not, of course, the same as the Irish term 'Sheela-na-gig'. Such examples were flagged in the thesaurus.

Explorer's authoring tools

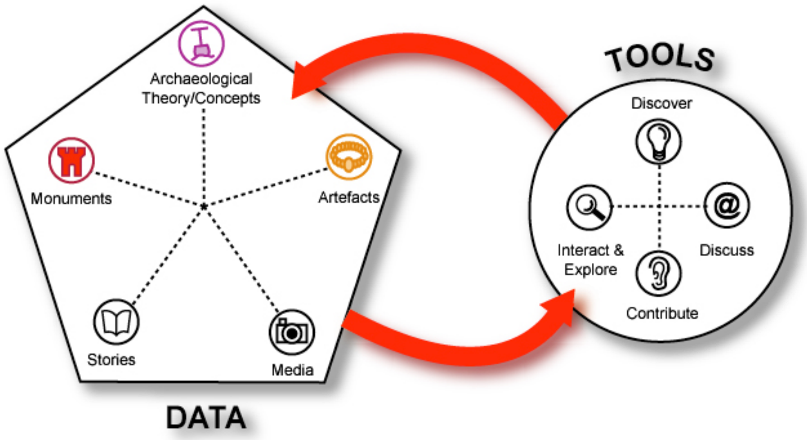


Figure 23: Depicts the relationship between the domain data and tools in Explorer forum.

Two authoring tools were developed to support the creation of narrative presentations within the forum: the *Story Creator* and the *Story Theme Creator*. The tools provide the user with a uniform environment for preparing narrative presentations and their associated annotations as illustrated in Figure 23. However, their functionality is closely associated with the CIPHER knowledge framework (Mulholland and Zdrahal 2003), principally:

- **Knowledge Acquisition:** the tools allow for the insertion and organisation of knowledge in terms of heritage artefacts, story characteristics, etc.

undertaken to relate the Irish monument types to terms within the EH ontology. The Irish classification system represents a resource of terminology which describes the built heritage of Ireland. Matching these terms with the EH thesaurus provided an interrelated tool for use on *Explorer* forum. In each case a determination was made as to the type of relationship which the class had with the EH thesaurus. However, the EH thesauri did not represent the concept of time as discussed with the Discovery Programme staff. It was therefore decided to create a controlled vocabulary to regulate the time periods used in the forum. This task was also undertaken by researchers from the Discovery Programme staff. Twenty six time periods, from pre-historic to modern, were identified as being necessary to represent the Irish chronology.

In order to use the English Heritage (EH) thesaurus of monument types, a mapping process had to be undertaken for each Irish monument type and archaeological object type to find where possible a matching terms within the EH thesaurus. The Irish classification system itself represents a resource of terminology which describes the built heritage unique to Ireland.

The mapping process therefore produced the following results. Of the 786 classes in the Irish vocabulary, 472 (60%) had a direct match to a term in the EH thesaurus. In this case the term was directly mapped. However 224 (28%) terms were closely related to terms in the EH ontology but were not linguistically similar enough to provide a direct mapping. In these cases the terms were mapped as preference terms, for example ‘Ring fort’. A further 101 terms (13%) were used principally in the context of Irish folklore or archaeology and did not have a match with any English heritage terms. These terms are viewed as candidates for new classes within the ontology. Examples of these terms are ‘Fian-bhoth’ and ‘Baulk’.

Figure 22 - Integration between presentation, business and integration tiers.

The Integration Tier

The central component of the integration layer is the Data Access Object (DAO) as illustrated in Figure 22. The DAO software pattern abstracts data retrieval between the resource and business layer. The benefits of incorporating the DAO pattern into the integration layer was evident as it was a potential requirement to port the forum application to another database vendor; as it would dramatically decrease development time by extracting vendor specific code from the business logic.

The Resource Tier

The resource tier was comprised of a set of MySQL databases: the narrative database, sites and monuments record (SMR), adapted English heritage thesaurus (both findings and monuments) and a media database. Integration amongst data sources was coordinated in the integration tier. Each resource was accessed using the MySQL connector/j driver[84]. The driver offers an extensive set of SQL functions for querying MySQL datasets.

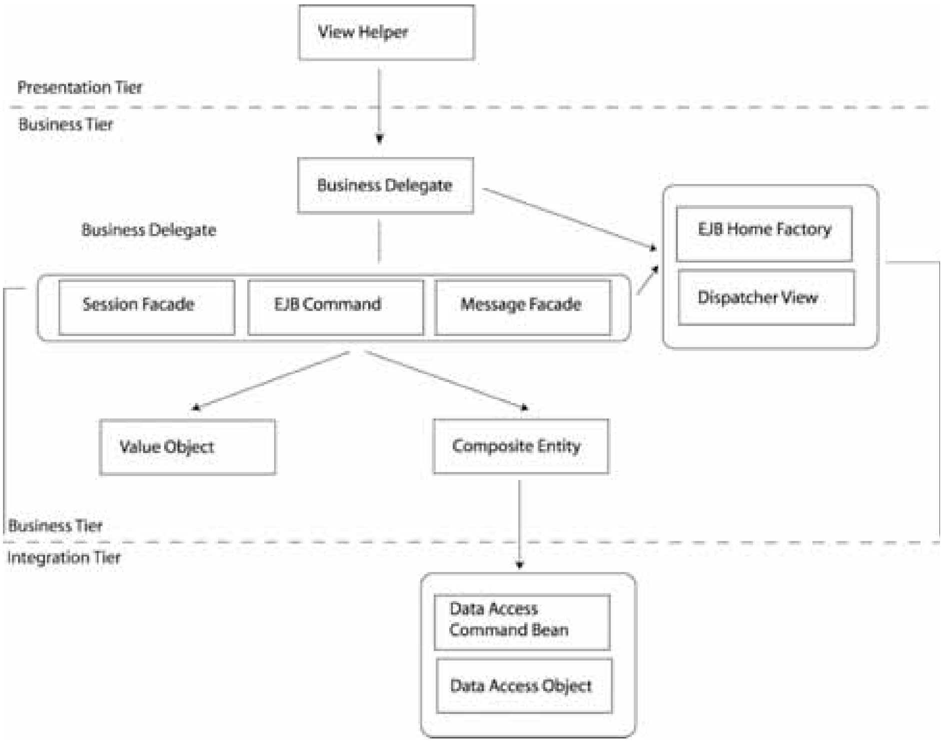
Modifying the English Heritage Thesauri

Domain experts examined each of the classcode terms in the Irish catalogue. There are a total of 786 classcodes in the Irish classification system. In order to use the English Heritage (EH) thesaurus of monument types, a matching process was

[84] For more information on the java implementation of the Mysql driver see http://www.mysql.com/products/connector/j/.

The business tier

The business tier supports the business logic of the application. This tier was implemented using the Enterprise Javabeans[83] framework, which provides distributed enterprise architecture for the development of real-world enterprise-wide applications. The EJB framework uses the 'Write Once Run Anywhere' (WORA) principle of early Java systems; however the architecture is not only platform independent but implementation independent. Within this context EJB applications can be ported to a number of platforms and application servers running a number of proprietary database servers. This feature highlights the flexibility supplied through the EJB architecture. Consulting the design pattern catalogue a number of patterns were recognised as appropriate for the development of the business tier as illustrated in Figure 22.



[83] http://java.sun.com/products/ejb/

The presentation tier

The presentation tier was developed using the Java servlet API (Hunter 1998) and the JSTL[82] (JavaServer Pages Standard Tag Library) presentation framework. The servlet API provides a rich functional language for the development of dynamic web applications while the JSTL framework supports the much of the core functionality of web applications through a simple tagging syntax similar to both HTML and XML. Using both technologies simultaneously decouples the development and design processes found in dynamic web site design. The developer may focus on functionality while the designer deals solely with presentation; this decoupling of functionality is an implementation of the early MVC, (Model View Controller) design pattern used widely for web development. Consulting the design pattern catalogue a number of patterns were recognised as appropriate for the development of the presentation tier of Explorer, they are the front controller, view helper, composite view the service to worker and the dispatcher view as illustrated in Figure 21.

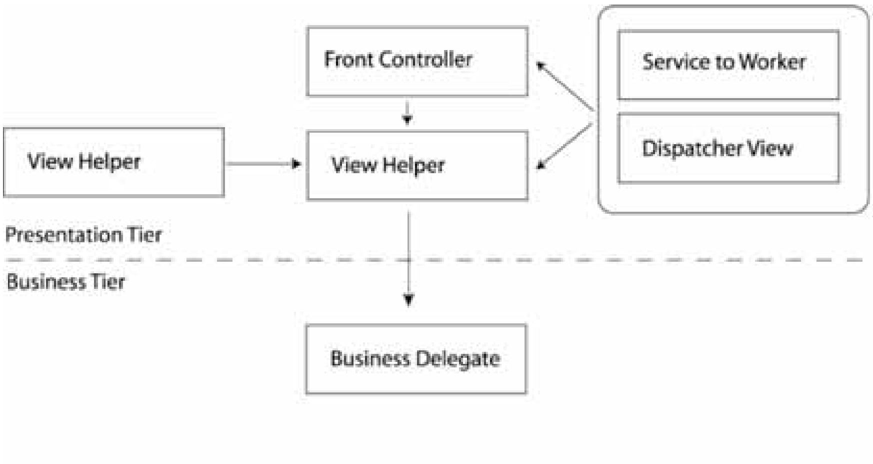


Figure 21 - Presentation tier design patterns

[82] For more information on JSTL see http://java.sun.com/products/jsp/jstl/.

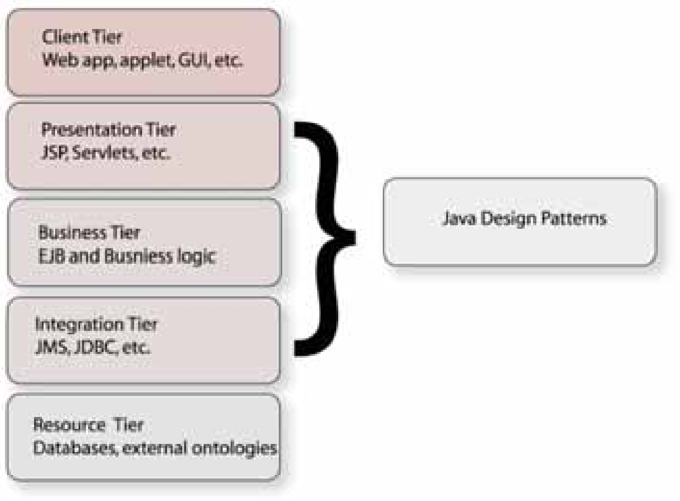


Figure 20 - Relation between tiered architecture and Java design patterns.

Two design pattern catalogues were created to accompany the development of the J2EE applications, namely sun's J2EE patterns catalogue[80] and the serverside.com's design patterns[81]. A design pattern is a proven solution to a problem in a context (Juric, Nashi et al. 2002) and provides a developer with a reusable model to solve recurring problems in a software engineering lifecycle. As illustrated in Figure 20, three tiers were identified as candidates for the application of J2EE design pattern, they are the presentation tier, business tier and integration tier. Design patterns are not definitive methodologies, yet they provide developers with tried and tested methods of problem solving. Within this context a number of design patterns were seen as applicable to the development of the Explorer forum. The following section describes the development of the presentation, business, integration and resource tiers through the of specific design patterns as described in (Juric, Nashi et al. 2002).

[80] The J2EE core design patterns are available at http://java.sun.com/blueprints/corej2eepatterns/index.html:

[81] The Serverside design patterns can be found at http://www.theserverside.com/patterns

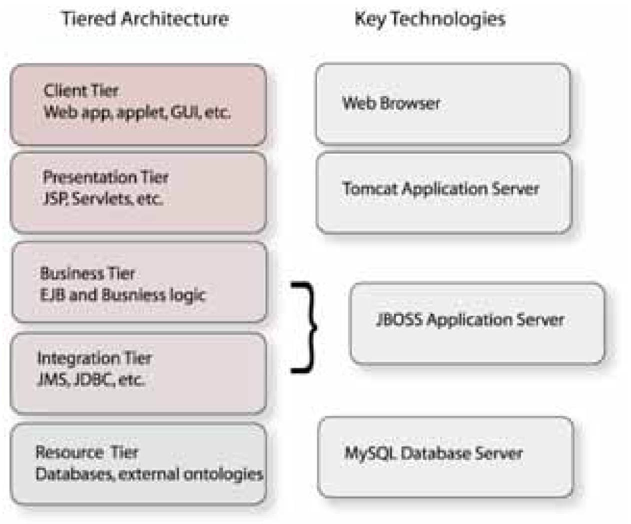


Figure 19 - Tiered architecture and accompanying technologies

The development cycle of the Explorer forum surrounds two separate trial dates. The first trial focuses on testing the CIPHER methodologies and the second on refining the forum's tools. As the forum is continually in development during this time frame, it is imperative to create well-structured modular software to allow for continual refinement in the interest of good design practice. A tiered approach is prescribed when developing with the Enterprise Java framework. This approach is an implementation of the MVC (Model View Controller) design pattern used to divide functionality into components for ease of design and software maintenance. A tier is a logical separation of concerns in a system. At this level of abstraction each tier is assigned a unique responsibility in the overall system allowing for tiers to be separated from other tiers while being loosely coupled to the adjacent tier. The tiered approach provides the developer with a lose coupling of modular functionality allowing for the addition of new tiers, for example data sources, without deconstructing the entire software framework.

OWL allows the restrictions to be placed on property types and there are of a number of inference engines currently available (Racer[78], Pellet[79]). However, the granularity of the ontology dictated the type of language chosen. It was agreed that a relational database provided ample functionality for the Explorer forum; furthermore, without a lengthy development period much of the functionality inherent in the use of ontologies will go unused. Within this context, MySQL was chosen for as the relational database server.

Although a brief analysis of database systems was carried out, MySQL was chosen due to it large user base and as it is widely accepted as the primary choice of a database server for the development of open source web applications. Its popularity is emphasised with its inclusion in the LAMP (Linux Apache MySQL PHP) framework which is gaining in popularity and rivalling more propriety solutions such as Java's Enterprise Edition and Microsoft's .Net framework.

Design and Implementation

[78] Racer can be downloaded from http://www.sts.tu-harburg.de/~r.f.moeller/racer/. Racer will work as a standalone server with the protégé ontology development environment.

[79] Pellet is an open source Java reasoning software is available for download at http://pellet.owldl.com/.

PHP	X	X	X	X	X
Python	X	X	X	X	X

Table 1 - Requirements and Properties of Software platform

Descriptive Language Choice

The descriptive power of the ontology is the key factor contributing to the choice of ontology language. The brief evaluation above describes three of the languages considered for the Explorer forum. All three are based on graph formalisms and are gaining in popularity due to the increased interest in knowledge based systems. However, the Explorer ontology consisted of two thesauri and a controlled vocabulary all of which could be simply represented in a relational database. The need for a more descriptive formalism was not required; however each was evaluated in the possibility of enhancing the conceptual representation of the forum. All three of the languages have query interfaces and numerous interfacing API's. However, both RDF and OWL allow for the encoding of more implicit types of knowledge which may be unearthed through the use of an inference engine. The RDF language was developed to accommodate the increasing interest in the semantic web; the language provides a more descriptive representation of web resources but relies on a rules engine (Jess[75], Algernon[76], Jena[77]) for enhanced functionality. OWL represents ontologies to the far right of McGuinness's spectrum (Figure 4) and is part of the description-logic family.

[75] The Jess rules engine is available at http://herzberg.ca.sandia.gov/jess/.

[76] The Algernon inference engine, along with several resources, is available at http://algernon-j.sourceforge.net/. Algernon can work as plug-in with Stanford's' Protégé ontology development environment. Protégé is available at http://protege.stanford.edu/ and supports ontology development in a variety of ontology languages.

[77] Jena is an open source semantic web framework available at http://jena.sourceforge.net/.

development of open source software. From this perspective, the partners aimed to make use of the growing collection of tools being developed by the open source community. Microsoft's .Net framework did not fulfil this requirement. Although the framework has gained many supporters from industry, and supports multiple languages through generic compiler architecture, it is a proprietary standard which includes a number of proprietary tools to support the software lifecycle.

Again .Net was one of only a few frameworks not to support other non proprietary operating systems such as the various flavours of Linux. The project partners aimed to develop tools that did not rely on any single operating system. As can be seen from Table 1, the other languages conformed to the set of project requirements put forward for the development of the Explorer forum. Within this context, each of the languages could be used for development. However, the development team's skill set is very important in choosing a software framework. The development team had previous experience with the Java language and therefore Java was chosen for the Explorer forum's development. The Java language is widely used to prototype applications during research and development. The language is portable, has a clean syntax, is interpreted and has a large range of API's on which to develop. Furthermore, the Java 2 Enterprise Edition is widely recognised within industry as a highly developed enterprise framework for the development of advanced web applications.

Language	CGI	Database	Web services	O/S	Platform
Java / J2EE	X	X	X	X	X
.Net	X	X	X	-	-

too rigid and for experimental research development. Although the initial functionality was specified, the model did not allow sufficient room for more innovative approaches to be applied during the lifecycle. It was therefore felt that the spiral lifecycle combined the best points from both the two previous models and in this respect was the most appropriate lifecycle for the development of the Explorer forum. The model required the creation of an early prototype as in the prototype model but with staggered implementation as with the waterfall model. This lifecycle suited the project's time frame as regards the user trial dates and offers a degree of freedom required for a research project. The initial stage of all three models is the evaluation of key technologies and the following describes the technologies chosen to support the Explorer forum.

Development Framework Choice

The Explorer forum has as an objective the contribution of cultural heritage narratives from an online community of interest. This requirement specifies the development of a dynamic web application. Current dynamic web applications are developed server side and use the functionality of HTML and the hypertext protocol to transmit between client and server. The client-server architecture provides for the bulk of processing to be carried out server-side while the client deals with presentation and simple processing. However, the architecture requires a server side language to exhibit dynamic functionality. A brief overview of some of the possible languages was presented above. However, not all of those languages exhibit all of the requirements set forward by the project. Firstly the language had to support the CGI architecture, have database connectivity and offer mature web service integration. Furthermore, from the outset, researchers at DIT were committed to supporting the

OWL DL: Is for users whose requirements include maximum expressiveness without losing computational completeness and decidability of reasoning systems. OWL DL supports all OWL constructs and has desirable computational properties for reasoning systems as with DL formalisms

OWL Full: Provides the user with maximum expressiveness with no computational guarantees.

Evaluation of key technologies

The following discusses the choice of development cycle, software framework and descriptive language with regards to development of the Explorer forum.

Lifecycle Model Choice

Researchers at DIT developed Explorer over a two and half year time frame. Within this period, specific milestones were set to help structure the project's lifecycle. The most applicable technologies for the lifecycle was determined by two separate user trial dates, the first to help the researchers test initial prototypes and the second to allow for refining of current software models. The introduction of two user trial dates heavily influenced the project's development lifecycle choices. The waterfall cycle, for example, required a long design period before the development of a prototype took place. However, the project required the development of an early prototype to test during the first trial. Therefore it was felt that the waterfall model was inappropriate.

The second lifecycle to be considered was the prototype model, which as the name suggests, proposes the development of an early prototype. Although the obvious advantage of an early prototype was apparent, it was felt that the lifecycle could be the resource and the subject is the value of the property. However, vocabularies are needed to indicate how the triples are supposed to be processed, specifically when describing specific kinds or classes of resources and the properties required in describing those resources. RDF itself provides no means for defining application-specific classes and properties, instead they are described using RDF schema. With the addition of RDF schema, RDF models can be viewed from an object-orientated perspective and thought of as frame-based knowledge representations. However, the language lacks the descriptive power required for advanced reasoning. The language is becoming increasing popular and is used in applications such as Mozilla's Firefox. There are a number of RDF (JENA, Seasame) stores available which provide a means for storing RDF graphs and APIs for simple reasoning and querying.

Ontology Web Language (OWL)

The Ontology Web Language (Smith, Welty et al. 2004) is a semantic mark-up language derived from DAML+OIL (Connolly, Harmelen et al. 2001) which is a based on description logic formalisms. The language is part of the semantic web initiative and is intended to provide web resources with additional semantics to aid software agents in processing and reasoning about web content. Additional semantics are supported by the instantiation of heterogeneous web ontologies. OWL builds on XML's ability to describe default tagging schemes and RDF's ability to formally describe the meaning of the terminology use in web documents.

RDF schema does not possess the descriptive power needed for reasoning. OWL builds on RDF to allow this reasoning but may be broken into three "flavours" designed for use by specific communities.

OWL Lite: Supports simple constraints and classification hierarchy.

Resources Definition Framework (RDF)

The Resources Definition Framework (RDF) (Lassila 1997; Lassila and Swick 1999) is a metadata specification for representing information about resources on the Web. RDF is an umbrella framework for representing information resources with commonly defined formal semantics. The framework allows for the automated processing of resources by supporting interoperability between applications without the loss of information. With the RDF model, knowledge is represented as directed labelled graphs where nodes represent concepts and arcs represent the relationship between concepts. The difference between semantic networks and RDF is that the nodes and arcs in RDF are labelled by URIs (Universal Resource Identifier) making the graph useable on the web, or 'webised' (a term coined by Tim Berners-Lee (Berners-Lee 2001)). The example below shows how the Dublin Core metadata standard is integrated with vCard, the electronic business card metadata standard through RDF (Iannella 2001).

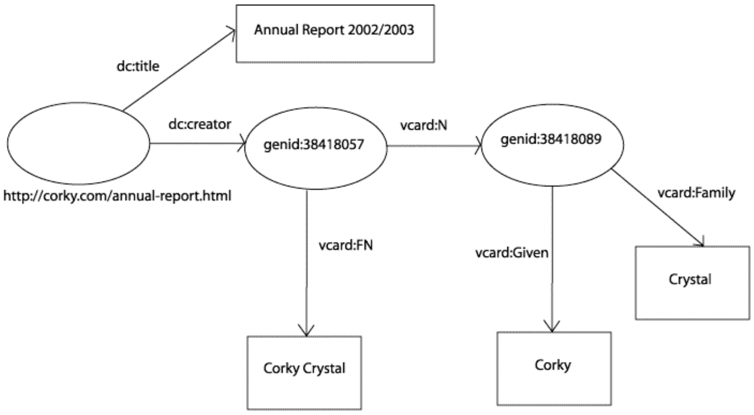


Figure 18 - RDF model of DC & Vcard

RDF provides a simple way of making statements about web resources. Each statement represents a triple consisting of an object, a predicate and a subject. The object is the resource being described, the predicate represents the property describing relationships (Pepper 2000). Topic maps provide a very simple way in which to model knowledge. The uptake and use of technology is often largely determined by its ease of use. The extensive acceptance of the WWW, it could be argued, was facilitated by the ease with which any person with rudimentary web publishing skills could obtain a presence on the web. This paradigm could be applied to the Topic Map architecture. However the standard does not contain the ability to conceptualise beyond the concept of a topic, therefore only offering the descriptive power of the topics associations. Each map is based on three core elements:

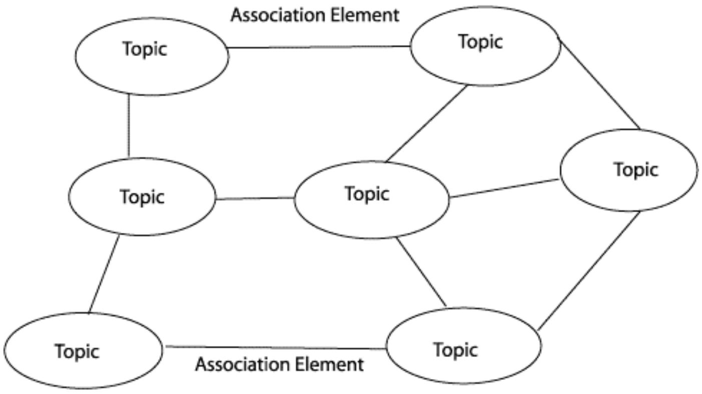


Figure 17 - Topic Map example

The Topic Element: The topic element is the building block of a Topic Map. A topic acts as a proxy for some subject while the subject reifies some real world object. The subject represents any idea, concept, document or information object. Example of such subjects would be a driver, the car and the 'driving' relationship.

The Association Element: The association element is a relationship between one or more topics.

Topic Occurrence: Is any information that is relevant to the subject represented by a topic. An occurrence is usually a resource that is addressable by a referencing URI. (Pepper and Moore 2001) modules, the presentation tier may be split between compiled code and scripting code making it easier to debug and script.

- Python is fully open source.

Summary of Semantic Web languages

With the advent of semantic web research new and improved means of distributing semantic information is being developed. Knowledge modelling began with the development of expert systems development during the 1980s. However, a renewed interest has sparked collaborative development of a number of representational languages to aid in the capturing and dissemination of semantic information. The following gives a brief overview of a number of representational languages considered during the evaluation stage of the CIPHER project's lifecycle. It is worth noting that this is by no means an extensive account of the current state of the art regarding ontology languages, however is does offer the reader an overview and relates to McGuiness semantic spectrum (Figure 4).

Topic Maps

Topic maps[74] were originally conceived as a method for indexing information resources. The ISO standard (ISO/IEC13250 2002) is related very closely to topic maps. The standard incorporates both conceptual graphs and semantic networks but by using XML as its foundation, provides a unique standard based way of encoding and exchanging knowledge. It offers a means for modelling abstract concepts and applying those abstractions to different information pools by identifying concept

[74] For more information on topic maps see http://www.topicmaps.org/.

the acronym PHP, however recently the language was endorsed as a key player within the enterprise development community as seen by its inclusion in the LAMP (Linux, Apache, MySQL, PHP) enterprise framework.

Python

The creator of Python, Guido van Rossum, describes Python as;

'*an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, makes it an ideal language for scripting and rapid application development in many areas on most platforms*'.

Like PHP, Python is a scripting language with the ability to embed dynamic tags in HTML pages but unlike PHP Python has a smaller developer community and user base. The following is a list of Python's features;

- Python facilitates fast development due to its easy syntax.
- It's runtime performance is well regarded though its use of C++ modules.
- Python allows developers to write their own C/C++ modules and load them into their Python code. Also Jython or the Java implementation of Python integrates Python with Java modules.
- Python supports XML and web services through SOAP.
- Python code is portable and will run on nearly every operating system available.
- Python is a scripting language and is not compiled. This can be a disadvantage in large scale applications, but with support for Java and C++

- Like Java and C#, it is interpreted.
- It supports Object Orientation and inheritance.
- PHP is open source. PHP is also well supported and itself supports the open source database MYSQL. There are also many open source debugging and development tools available.
- PHP has good XML and SOAP support for web services.
- It supports Java by providing Servlet execution and instantiation and manipulation of Java classes as simple PHP classes.
- PHP has many open source modules available which rapidly reduces development time for many applications and it has the support of a large open source development community.
- PHP will integrate with almost any database and will run on almost every platform.
- PHP uses simple syntax and is designed specifically for HTML scripting in contrast with other similar languages, such as Perl.
- It has the potential to reduce development and maintenance costs dramatically.
- PHP has the potential disadvantage of being un-compiled. This can be a disadvantage for large projects requiring scalability.

PHP has emerged as a well supported key technology due to the factors stated above. Its main advantage is development speed. PHP can be prototyped rapidly, is very robust and due to its clear syntax is easily maintained. The language was originally conceived as a means for developing dynamic personal home pages hence

- The CLR provides security features which are comparable to Java's Sandbox approach.
- While .Net framework is largely language agnostic, Microsoft brought out a Java/C type language called C# optimised for the framework. This has many of the key features of Java, such as garbage collection.
- The .Net framework itself is not open source and the tools needed to develop and implement the framework are created by Microsoft.
- .NET uses a number of open standards and has been developed with Web Services in mind from the outset. The CLR itself has been open to some standardisation of late.

In summary .Net is an enterprise solution which will compete with Java for the enterprise market in Web Services. Sun has responded by creating the SunOne rapid application development environment. The equivalent development environment from Microsoft is Visual Studio .NET. The battle for the enterprise market has intensified by the inclusion of a completely open source solution based on the scripting language PHP.

PHP

PHP is an open source html embedded language, which supports dynamic content on the web. PHP has become an important tool in dynamic web applications in use with such major organisations as YAHOO. PHP is seen as a useful method to rapidly develop a dynamic websites.

- PHP is tightly coupled with HTTP and HTML, as it is an embedded, scripting language.

led to the development of three separate APIs representing these three application areas; the standard edition provides for desktop applications, the micro edition for portable devices and the enterprise edition for enterprise wide and web applications. As the Explorer forum aimed to support a potentially large community of online users the enterprise edition was the most applicable and provided a comprehensive range of APIs. Furthermore, as the technology has matured, numerous frameworks have emerged to aid in the development process. The Servlet specification is a Java implementation of the common gateway interface (CGI) created to support Java applications interface with web clients. When coupled with the JavaServer pages specification the framework provides a powerful means for developing sustainable web applications while providing Java's inherent portability and object orientated functionality. A further implementation of the enterprise paradigm is the Enterprise JavaBeans framework, developed to provide increase distributed and database functionality. The J2EE architecture has been embraced within industry circles as a sustainable way of supporting distributed enterprise wide applications.

.NET

Microsoft has created the .NET framework to address enterprise wide application development and to an extent match the development segment J2EE was filling. The following points can be made about the Framework;

- Language neutrality: a Common Language Runtime (CLR) software component similar to the Java Virtual Machine (JVM) allows many languages to be used within the framework. For example C++ could seamlessly use Java methods and classes as they are all compiled into an Interpretive Language (IL) similar to Java byte code.

terminated with a customer evaluation. The spiral lifecycle can highlight potentially dangerous risk areas, however this process does rely heavily on the developer team.

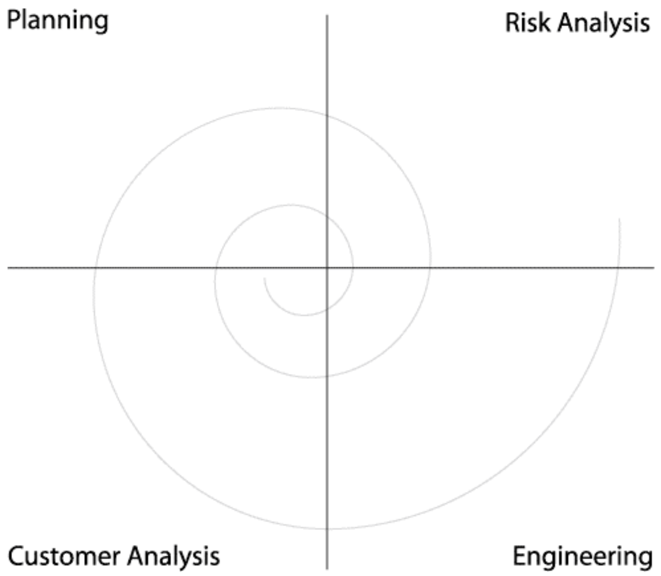


Figure 16 - Spiral model

Summary of web application languages

Web applications are generally developed using server side scripting languages or server side web frameworks, such as J2EE or Microsoft's .Net. Here the author provides a brief summary of the key points relating to examples of these technologies. This is by no means a comprehensive review but an attempt to summarise the difficulties and possibilities for developing complex integrated web applications at the time of the CIPHER project.

Java and J2EE (Java 2 Enterprise Edition) overview

The Java language has become a widely recognised development framework since its introduction in 1995. Initially heralded as the language of the internet, Java has grown into a complete software framework encompassing desktop programs, enterprise solutions and mobile phone applications. The additional functionality has iterations of the first two steps in refining the initial prototype. Once the customer is satisfied, development of the actual product begins. The prototype model is very useful when the customer does not have a clear understanding of the requirements. Consequently the requirements tend to be more stable and the risk of new development is reduced.

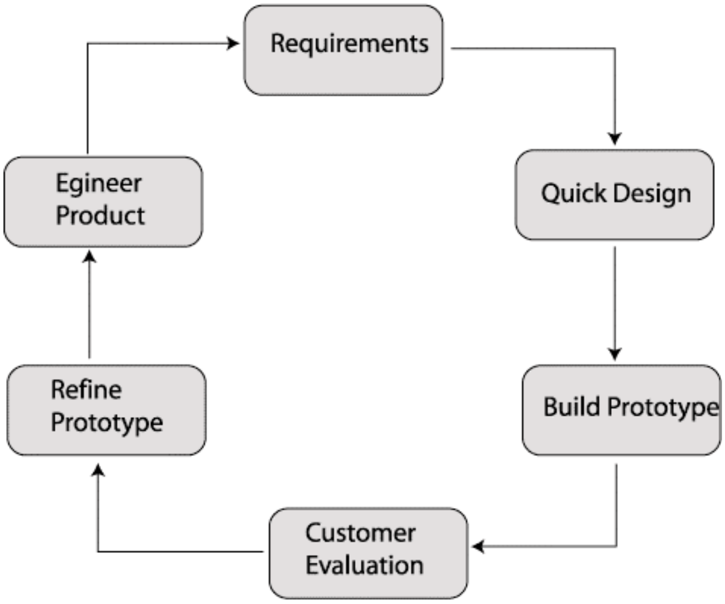


Figure 15 - Prototype model

Spiral Model

The spiral model is a hybrid representation of both the Prototype and Waterfall models. The approach consists of four key phases, Planning, Risk Analysis, Engineering, and Customer Evaluation. The spiral model combines the staggered implementation of the waterfall model with the prototyping of the prototype model. The model works by progressively building more complete versions of the software. Each complete loop of the spiral moves though the four quadrants as can be seen in Figure 16. A complete loop is initiated by a planning phase; followed by a risk analysis in which a project may be terminated if the risk of continuing seems too great. An engineering process phase develops the current prototype and the cycle is

- Development phase - implementing the concept and definition phases
- Evaluation phase - evaluating the result
- Operation phase - placing the software into use

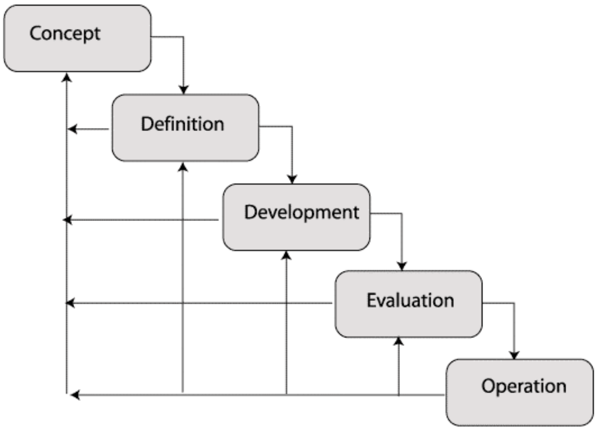


Figure 14 - Waterfall model

As can be seen in Figure 14, the waterfall model accurately reflects the lifecycle of a software project; it is iterative and may go through a single phase numerous times before completion. Although the waterfall model is the most common and popular of design methodologies, many believe the approach to be flawed as a working prototype is not produced until late in the development cycle. This belief has spawned a number of similar software lifecycles, based on the original waterfall model.

Prototype model

The prototype model (see Figure 15) was developed to tackle the problems associated with late prototyping inherent with the waterfall model. The cycle involves the gathering of initial requirements and then the rapid implementation of a prototype. The customer, in the case of the Explorer forum this was the user community, can then evaluate the prototype. This evaluation can involve further

Appendix B: Developing Explorer

The following presents a brief overview of the approaches and techniques employed during the development of the Explorer forum. The discussion concentrates on firstly presenting a summary of types of engineering cycle and relevant technologies, and secondly evaluating each according to the criteria of the forum and the skills of the development team. Next the discussion presents the design and implementation of the forum, describing each software tier in detail. Finally, some the Explorer forum tools are briefly explained.

Summary of software engineering lifecycles

The aim of this section is to offer the reader a brief overview of selected software lifecycles that were considered in supporting the development of the Explorer forum. The purpose is to demonstrate the different approaches available to software engineers and not to offer the reader a complete and exhaustive description of the software development process. Within this context, the overview focuses on three common engineering methodologies, the waterfall model, the prototype model and the spiral model.

The waterfall model

The 'waterfall model' is the simplest of the software engineering lifecycles. The model is broken into five distinct parts as displayed in, and can be briefly described as:

- Concept phase - establishes what the software should do
- Definition phase – establishes how the software will work

number of clearly defined steps. Shneiderman draws on both theories in creating his genex framework. Genex (named after Bush's memex mentioned) consists of four stages, the first of which echoes the model put forward by the *wunderkammer*; they are collect (learn from previous works), relate (i.e. consult with peers and mentors); create (explore, compose, and evaluate possible solutions); and donate (disseminate the results). Shneiderman approaches creativity from a human computer interaction (HCI) perspective. He outlines the challenges of developing software to support creativity and aims to approach the problem through the implementation of the genex framework. This is a narrower yet more relevant approach to creativity in terms of this thesis; however, the method in which the creative process is implemented is crucial to its proliferation.

often choose exhibits from the museum's reserve collections. Eduardo Paolizzi, when guest-curating in the British Museum, selected over two hundred objects from the reserve collection exhibiting a montage of items that would never usually be juxtaposed (Putnam 2001). His collections had parallels with the *wunderkammer*, or curiosity cabinet, where multiple associations simulate both thought and wonder.

The curiosity cabinet or *Wunderkammer* which was popular in Europe from sixteenth to the eighteenth century could be described as a forerunner of the modern museum. It consisted of a diverse compilation of unrelated objects collected to inspire creativity and curiosity. The concepts put forward by the *Wunderkammer* are recognised by more contemporary sources as a valid description of the creative process. Initial analysis of creativity began with Sigmund Freud and Carl Jung early in the 19th century. Their work helped to shift the focus from product to person when thinking of creativity. However, it was Wallas who introduced the concept of creativity as a process and in that helped to unveil the surrounding mystery. Now creativity is often defined as a serious of steps each lending to the next (Herbjørnsen 2003). Wallas' introduced a four stage model of creativity, preparation, incubation, illumination and verification.

Shneiderman, in his article Creating Creativity (Shneiderman 1999), describes other perspectives of creativity. Structuralist theorists, for example, have emphasised a more formal approach to writing creatively stressing the importance of studying previous literature and methodically approaching the problem. Alternatively, *situationlists* highlight the social side of creativity and approach the creative process in terms of a social progression embedded in a community of practice (similar to both Lave and Wenger theory of LPP (3.2) and Vygotsky's theory of social constructivism). Both approaches reflect Wallas' initial model of creativity through a environment they are transformed in the way they provide meaning. Examples of this are religious paintings which, when taken out of a cathedral or church and placed into the neutral surroundings of a museum, become artefacts to be studied and lose much of their previous significance. A visitor will, in this case, gather a different meaning when the piece is viewed out of context. It is generally recognised that learning and understanding is most effective when properly contextualised; therefore, visitors should ideally be allowed to explore collections of artefacts in context (Mulholland and Collins 2002). From this perspective, digital narrative has been proposed as an approach to re-establish the original context in which the artefact was found and how it came into being (McAuley and Carswell 2007; McAuley and Carswell 2008). This is because museums by their very nature often repress the process by which the objects are created.

Contemporary artists have tried to incorporate this process of creation into the exhibit thereby revealing the domain as a more dynamic and ongoing progression. Dion attempts to build in this method with his "*Tate Thames dig*" exhibition. This exhibition was essentially in two parts, the first exhibiting the process of the dig and the second exhibiting the excavated artefacts. Visitors were able to see newly excavated artefacts in tents outside the Tate Modern gallery, allowing patrons to experience the process of excavation. Inside the Tate, selected objects were displayed in vitrines (Putnam 2001). While museums collect for exhibitions, artists themselves support collecting as a creative and learning process. Two reasons are proposed why collecting inspires creativity, firstly because collecting has purpose and secondly because collecting has to contend with a set of constraints. The purpose might be a topic that interests the collector and the constraints might include the availability of the required objects (Mulholland and Collins 2002). When guest-curating, artists museum are presented with the curator's summary of the collection - not necessarily how the artist may have originally intended or imagined the presentation of the collected-work. Museum director Ned Rifkin compared this approach to that of a word processor, prompting the user with two options 'save and display' (Stein 2003). The ways in which museums traditionally present collections in a narrative sense have paved the way for introspective examinations on how museums conserve and exhibit. Both Barthes and Foucault maintained that museum collections should reflect the assumptions of the prevailing culture rather than presenting a curator's view of the subject (Putnam 2001). Contemporary artists have also shifted the view of curatorship from the traditional taxonomy to reflect more current theories of learning and narrative and in doing so often highlight the museum's earlier approach to exhibition. One of the most famous examples is of Fred Wilson's '*Mining the Museum'* exhibited in Maryland, USA in 1992. He sought to raise awareness of institutionalised racism, making explicit the subtle and insidious ways in which museum may select artefacts for exhibition (Stein 2003). Through the juxtaposition of archived artefacts with exhibited artefacts, Wilson aimed to highlight the fact that the museum collection is constructed from a particular perspective and it should be the visitor who actively interprets and reconstructs his experience from the collection. Wilson's work shows how the juxtaposition of artefacts can alter a visitor's interpretation, highlighting the assumptions made within the museum's own narrative and helping to support active interpretation of the collection.

Museums have often been criticised for repressing the context surrounding an exhibit. Karp & Levine recognised the difficulty this problem presents by stating that '*almost nothing displayed in museums was made to be seen in them*' (Vogel 1991). When objects are removed from their intended context and placed in a new

Appendix A: Narrative, collecting and creativity

There is a strong narrative influence on the activities of the museum curator. Often the curator will use an overarching narrative to present a specific collection. In this context, Pearce maintains that the 'organisation or creation of objects follow rules akin to those followed when constructing natural language' (Mulholland and Collins 2002). The artefacts or objects represent the lexicon and the way in which these objects are arranged in a collected works represents the syntax or grammar. Typically a museum curator will use artefacts to help illustrate their narrative. Lisa Roberts mirrors this view maintaining that exhibits place objects in a narrative perspective, explaining how the object is part of a larger story (Roberts 1997).

Despite collections themselves being intrinsically linked with narrative, individual items can be used to validate a narrative as proof or evidence in what Pearse describes as 'the power of the real thing' (Mulholland and Collins 2002). It is the object, which draws the visitor's focus and provides a more intense and satisfying experience, while additional narrative can offer another perspective or explanation not explicitly stated by the object itself. The Shroud of Turin, for example, may be associated with a suitable narrative explaining that the artefact may be a centuries-old fake. Without the additional perspective, the visitor may be unaware of the shroud's unique history. The shroud of Turin represents a good example of an object that in itself is only a part of a compelling and interesting story.

Because galleries and museums have a finite exhibition space and, in general, more exhibits than they can present publicly, presentation of an exhibition involves interpretation and judgement on behalf of the curator. Essentially, patrons of a process does not solely identify creating or developing ontologies from scratch. Often the process can involve taking an existing ontology and later refining it or indeed using some from of machine learning to develop a semantic model.

Self-Organising Maps (SOM): SOM, developed by Teuvo Kohonen, are a means of automatically arranging statistical data so that similar inputs are mapped according to their underlying semantics.

Online Community (OC): An online community is often a geographically disparate community of users who come together, communicate and exist in a virtual capacity.

Community of Practice (CoP): A community of practice is often regarded as a learning community comprised of professionals who organise themselves into homogeneous communities, learning knowledge from other members within the group.

Community of Interest (CoI): A community of interest is often comprised of individuals from a range of disciplines who congregate out of a common interest. In this context communities of interest are heterogeneous communities.

Artificial Intelligence (AI): Artificial Intelligence is the term giving to the group of practices and technologies by which machines are developed in a certain way to exhibit some form of intelligence.

Glossary

Knowledge Management (KM): Knowledge management describes the range of approaches, methodologies and supporting technologies used to collect, represent and employ real-world knowledge.

Knowledge Acquisition: In the context of this thesis, knowledge acquisition involves the collection and specification of knowledge for knowledge-based systems.

Knowledge Engineering: Knowledge engineering describes the process by which harvested or collected knowledge is encoded and specified in some form of knowledge model.

Knowledge Representation: Knowledge representation, in the context of this thesis, is the broad-term used to describe a multiple of ways in which knowledge is captured, codified and represented for re-use.

Simple Ontology: A simple ontology is a type of knowledge model that is not encoded in a machine-readable format, and is, therefore, more suited for use by human users. Unlike formal ontologies, simpler ontologies do not attempt to represent stocks of tacit knowledge. Generally simple ontologies provide a means to browse/search a repository with additional functionality such as disambiguation. They include controlled vocabularies, glossaries and thesauri.

Structured Ontology: A structured ontology attempts to represent both explicit and implicit knowledge in a machine-readable format. Human users mostly develop structured ontologies for use by sophisticated reasoning or inference software.

Ontology Acquisition: Ontology acquisition refers to the practice by which ontologies are acquired. This is different from ontology engineering because the

Wolfe, M. (2000). "Metadata, Knowledge Management, and Communications." Canadian Journal of Communication [Online] **25**(4).

Srinivasan, R. and J. Huang (2005). "Fluid ontologies for digital museums." Int. J. Digit. Libr. **5**(3): 193-204.

Stanoevska-Slabeva, K. and B. Schmid (2001). A Typology of Online Communities and Community Supporting Platforms. Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 7 - Volume 7, IEEE Computer Society.

Stein, J. E. (2003). "Sins of Omission: Fred Wilson's Mining the Museum." Retrieved 22 May 2005s, from http://slought.org/files/downloads/publications/salons/1083.pdf.

Sure, Y., H. Akkermans, et al. (2003). On-To-Knowledge: Semantic Web Enabled Knowledge Management. Web Intelligence. J. L. Ning Zhong, Yiyu Yao, springer**:** 30.

Surowiecki, J. (2004). The Wisdom of Crowds : Why the smarter are smarter than the few, ABACUS.

Taleb, N. N. (2007). The apprenticeship of an Empircal skeptic The Black Swan: The Impact of the Highly Improbable, Random House**:** 15.

Taleb, N. N. (2007). From Mediocristan to Extremistan, and Back. The Black Swan: The Impact of the Highly Improbable, Random House**:** 223.

Taleb, N. N. (2007). The scandal of predication. The Black Swan: The Impact of the Highly Improbable, Random House**:** 146.

Turoff, M. and S. R. Hiltz (1996). Computer Based DELPHI Processes. Gazing into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health E. Ziglio. London, Kingsley Publishers.

Uschold, M. and M. Gruninger. (1996). "Ontologies, Principles, Methods and Applications." Retrieved 4 August 2003, from http://citeseer.ist.psu.edu/cache/papers/cs/3214/http:zSzzSzwww.cm.cf.ac.ukzSzUserzSzJ-C.PazzagliazSzReferenceszSzart:Uschold-96.pdf/uschold96ontologie.pdf.

Uschold, M. and M. King (1995). "Towards a Methodology for Building Ontologies."

Vogel, S. (1991). Chapter12: Always True to the Object, in Our fashion. Exhibiting Cultures: The Poetics and Politics of Museum Display. I. Karp and S. D. Lavine, Smithsonian Books (June 1991)**:** p191.

Voss, J. (2007). Tagging, Folksonomy & Co - Renaissance of Manual Indexing? 10thInternational SymposiumforInformationScience. Cologne.

W3C (2004). World Wide Web Consortium Issues RDF and OWL Recommendations: Semantic Web emerges as commercial-grade infrastructure for sharing data on the Web.

Welty, C. (1999). Ontologies: Expert Systems all over again? American Association for Artificial Intelligence National Conference, Orlando, Florida, U.S.A.

Wenger, E., R. McDermott, et al. (2002). Communities of Practice and Their Value to Organisation. Cultivating Communities of Practice, Harvard Business School Press**:** 4.

Putnam, J. (2001). The Museum Affect. Art and Artifact - The Museum as Medium, Thames & Hudson**:** 40.

Putnam, J. (2001). Open the box. Art and Artifact - The Museum as Medium, Thames & Hudson**:** 26.

Redfern, T. and E. Kilfeather (2004). A Method for Presenting High Resolution, Archaeological 3D Scan Data in a Narrative Context. Proceedings of the Database and Expert Systems Applications, 15th International Workshop on (DEXA'04) - Volume 00, IEEE Computer Society.

Rheingold, H. (1998). The Virtual Community, The MIT Press.

Roberts, L. C. (1997). From Knowledge to Narrative: Educators and the Changing Museum, Smithsonian Books (June 1, 1997).

Rosenfeld, L. and P. Morville (1998). Information Architecture for the world wide web, Oreilly.

Sanger, L. (2006). Toward a New Compendium of Knowledge

Schank, R. C. (1995). What We Learn When We Learn by Doing. Chicago, Illinois, USA, Institute for the Learning Sciences Northwestern University.

Shirky, C. (2005) "Ontology is Overrated: Categories, Links, and Tags." Clay Shirky's Writings About the Internet **Volume**, DOI:

Shirky, C. (2008). Personal Motivation meets Collaborative Production. Here comes Everybody, Allen Lane**:** 122.

Shirky, C. (2008). Personal Motivation meets Collaborative Production. Here comes Everybody, Allen Lane**:** 109 - 120.

Shirky, C. (2008). Personal Motivation meets Collaborative Production. Here comes Everybody, Allen Lane**:** 127 - 130.

Shirky, C. (2008). Promise, Tool, Bargain. Here comes Everybody, Allen Lane**:** 261 - 292.

Shirky, C. (2008). Sharing Anchors Community. Here comes Everybody, Allen Lane**:** 27.

Shirky, C. (2008). Sharing Anchors Community. Here comes Everybody, Allen Lane**:** 48.

Shneiderman, B. (1999). Creating Creativity for Everyone: User Interfaces for Supporting Innovation. Maryland, (umiacs) University of Maryland Institute for advanced computer studies.

Sinha, R. (2005). A cognitive analysis of tagging (or how the lower cognitive cost of tagging makes it popular).

Smith, M. K., C. Welty, et al. (2004). OWL Web Ontology Language Guide, W3C.

Sowa, J. (2000). Ontology, Metadata, and Semiotics. International Conference on Conceptual Structures, ICCS'2000, Darmstadt, Germany.

Srinivasan, R. (2003). Village Voice: An Information-based Architecture for Community-centered Exhibits. Museums and the Web, Toronto Ontario.

McGuinness, D. L. and F. v. Harmelen (2004). OWL Web Ontology Language.

Minsky, M. (1975). A Framework for Representing Knowledge. The Psychology of Computer Vision.

Mulholland, P. and T. Collins (2002). Using Digital Narratives to Support the Collaborative Learning and Exploration of CH. IEEE International workshop on Presenting and Exploring Heritage on the Web (PEH'02) in conjunction with the 13th International Conference and Workshop on Database and Expert Systems Applications (DEXA 2002), Aix-En-Provence, France. .

Mulholland, P., T. Collins, et al. (2004). Story Fountain: Intelligent support for story research and exploration. Intelligent User Interfaces (IUI'2004), Madeira, Portugal.

Mulholland, P. and Z. Zdrahal (2003). Knowledge Support for Story Construction, Exploration and Personalisation in Cultural Heritage Forums. 14th International Conference and Workshop on Database and Expert Systems Applications (DEXA 2003), Prague, Czech Republic.

Mulholland, P., Z. Zdrahal, et al. (2002). CIPHER: Enabling Communities of Interest to Promote Heritage of European Regions. Cultivate Interactive.

Murray, J. H. (1997). Agency. Hamlet on the Hollodeck: The future of narrative in cyberspace. Cambridge, The MIT Press**:** 152.

Murray, J. H. (1997). Immersion. Hamlet on the Hollodeck: The future of narrative in cyberspace. Cambridge, The MIT Press**:** 98.

Noy, N. F. and D. L. McGuinness (2002). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford CA., Stanford Medical Informatics.

Noy, N. F. and M. A. Musen (1999). SMART: Automated Support for Ontology Merging and Alignment. Twelfth Workshop on Knowledge Acquisition, Modeling, and Management, Banff, Canada.

O'Reilly, T. (2005) "What Is Web 2.0." **Volume**, DOI:

Orr, J. E. (1996). Talking About Machines: Ethnography of a Modern Job (Collection on Technology & Work), Cornell University Press.

Pepper, S. (2000). "The TAO of Topic Maps

Finding the Way in the Age of Infoglut." Retrieved 27 June 2002, from http://www.ontopia.net/topicmaps/materials/tao.html.

Pepper, S. and G. Moore. (2001). "XML Topic Maps (XTM) 1.0." Retrieved 27 June 2002, from http://www.topicmaps.org/xtm/1.0/#ref_iso13250.

Preece, J. (2000). Chapter 1 - Introduction. Online communities: Designing usability, supporting sociability. Chichester, John Wiley & Sons**:** 19.

Preece, J. and D. Maloney-Krichmar (2003). Online Communities: Focusing on sociability and usability Handbook of Human-Computer Interaction,. Jacko J. & Sears A. NJ, Lawrence Erlbaum Associates Inc.

Preece, J., D. Maloney-Krichmar, et al. (2003). History and emergence of online communities. Encyclopedia of Community: From Village to Virtual World. K. Christensen and D. Levinson, SAGE Publications**:** 2000.

Landow, G. P. (1991). Roland Barthes and the Writerly Text. Hypertext: The Convergence of Contemporary Critical Theory and Technology, Johns Hopkins Univ Press: 5-6.

Lassila, O. (1997). Introduction to RDF Metadata. Cambridge (MA), World Wide Web Consortium.

Lassila, O. and R. R. Swick (1999). Resource Description Framework (RDF) Model and Syntax Specification. Cambridge (MA), World Wide Web Consortium.

Lave, J. and E. Wenger (1991). Situated Learning : Legitimate Peripheral Participation (Learning in Doing: Social, Cognitive & Computational Perspectives). Situated Learning : Legitimate Peripheral Participation (Learning in Doing: Social, Cognitive & Computational Perspectives). R. Pea, J. S. Brown and C. Heath, Cambridge University Press: 29.

Leiner, B. M., V. G. Cerf, et al. (2003). A Brief History of the Internet. Histories of the Internet, The Internet Society.

Lodge, D. (1990). Chapter 9 - Narration with words. Images and understanding : thoughts about images, ideas about understanding. H. Barlow, C. Blakemore and M. Weston-Smith. Cambridge, New York, USA., Cambridge University Press: 141 - 154.

Maedche, A. and S. Staab (2001). "Ontology Learning for the Semantic Web." IEEE Intelligent Systems **16**(2): 72-79.

Manola, F. and E. Miller (2004). RDF Primer. B. McBride, W3C.

Mateas, M. and P. Sengers (1999). Narrative Intelligence. Narrative Intelligence Symposium AAAI 1999 Fall Symposium Series, North Falmouth, Massachusetts.

Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Computer Mediated Communication - LIS590CMC, University of Illinois Urbana-Champaign.

McAuley, J. and J. Carswell (2007). An Open Approach to Contextualising Heterogeneous Cultural Heritage Datasets. 35th Annual Conference on Computer Applications and quantitative methods in Archaeology (CAA 2007). Berlin, Germany.

McAuley, J. and J. Carswell (2008). Knowledge Management for Disparate Etruscan Cultural Heritage. Second International Conference on the Digital Society (ICDS 2008). Sainte Luce, Martinique.

McAuley, J. and E. Kilfeather (2005). A comparative analysis of ontological techniques in supporting online communities. Computer Applications and quantitative methods in Archaeology CAA, Tomar, Portugal.

McCarthy, J. (1984). Some expert systems need common sense. Proc. of a symposium on Computer culture: the scientific, intellectual, and social impact of the computer. New York, United States, New York Academy of Sciences.

McCarthy, J. (2007) "What is Artificial Intelligence?" **Volume**, DOI:

McGuinness, D. L. (2002). Ontologies Come of Age. Spinning the Semantic Web. D. Fensel, J. Hendler, H. Lieberman and W. Wahlster, The MIT Press: 392.

Honkela T. et al. (1996). Newsgroup Exploration with WEBSOM Method and Browsing Interface. Helsinki, Lab. of Computer and Information Science. @ Helsinki University of Technology.

Horridge, M., H. Knublauch, et al. (2004). A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools

The University Of Manchester.

Hung, D. and M. Nichani (2002). "Differentiating between communities of practice (CoPs) and quasi-communities: can CoPs exist online." International Journal on E-Learning **1**(3): 23-29.

Hunter, J. (1998). Java Servlet Programming, O'Reilly.

Iannella, R. (2001). Representing vCard Objects in RDF/XML, W3C.

ISO2788 (1986). Guidelines for the establishment and development of monolingual thesauri, International Organization for Standardization (ISO).

ISO5964 (1985). Guidelines for the establishment and development of multilingual thesauri, International Organization for Standardization (ISO).

ISO/IEC13250 (2002). Topic Maps.

Jacob, E. K. (2004). Classification and Categorization:A Difference that Makes a Difference.

Juric, M., N. Nashi, et al. (2002). Introduction. J2EE Design Patterns Applied, Wrox**:** 9.

Kilfeather, E., J. McAuley, et al. (2003). Cultural Heritage Forum Desgin. International Conference on Hypermedia and Interactivity in Museums (ICHIM), Paris.

Kilfeather, E., J. McAuley, et al. (2003). An ontological application for archaeological narratives. 14th International Workshop on Database and Expert Systems Applications (DEXA'03), Prague, Czech Republic.

Kim, A. J. (2000). PURPOSE: The Heart of Your Community. Community Building on the Web : Secret Strategies for Successful Online Communities, Peachpit Press**:** 380.

Kohonen T. (1982). "Self-organizing formation of topologically correct feature maps." Biological Cybernetics **45**(5969).

Kollock, P. (1996). Design Principles for Online Communities. Harvard Conference on the Internet and Society. Boston, U.S.A.

Landauer, T. K., P. W. Foltz, et al. (1998). "Introduction to Latent Semantic Analysis." Discourse Processes **25**: 259 - 284.

Landow, G. P. (1991). The Definition of Hypertext and Its History as a Concept. Hypertext: The Convergence of Contemporary Critical Theory and Technology, Johns Hopkins Univ Press**:** 3-4.

Landow, G. P. (1991). Hypertext: The Convergence of Contemporary Critical Theory and Technology, Johns Hopkins Univ Press.

Fensel, D. (2000). Ontologies. Ontologies: a Silver Bullet for Knowledge Management and Electronic Commerce, Springer**:** 12.

Fernandez, M., A. Gomez-Perez, et al. (1997). METHONTOLOGY: from Ontological Art towards Ontological Engineering. The AAAI97 Spring Symposium Series on Ontological Engineering, Stanford, USA.

Fischer, G. (2001). Communities of Interest: Learning through the Interaction of Multiple Knowledge Systems. 24th IRIS Conference, Bergen, Norway.

Flexer A. (1999). On the use of Self-Organising maps for clustering and visualisation.

Frauenfelder, M. (2004). Sir Tim Berners-Lee: Tim Berners-Lee invented the World Wide Web, but he had something bigger in mind all along. He tells TR how his 15 years of work on the "Semantic Web" are finally paying off. Technology Review published by MIT. Boston.

Frey, J. G., G. V. Hughes, et al. (2003.). Less is More: Lightweight Ontologies and User Interfaces for Smart Labs. UK e- Science All Hands Meeting, Nottingham.

Friesen, N. (2002). "Semantic Interoperability and Communities of Practice." Retrieved 23 May 2004, from http://www.cancore.ca/documents/semantic.html.

Gamper, J., W. Nejdl, et al. (1999). Combining ontologies and terminologies in information systems. International Congress on Terminology and Knowledge Engineering, Innsbruck, Austria.

Gruber, T. (1993). "What is an Ontology?" Retrieved 21 October 2002, from http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.

Gruber, T. (2007). "Ontology of Folksonomy: A Mash-up of Apples and Oranges." International Journal on Semantic Web & Information Systems **3**(2).

Guarino, N. (1998). Formal Ontology and Information Systems. Formal Ontology in Information Systems, Trento.

Halpin, H., V. Robu, et al. (2007). The complex dynamics of collaborative tagging. Proceedings of the 16th international conference on World Wide Web. Banff, Alberta, Canada, ACM.

Hein, G. E. (1991). Constructivist Learning Theory. CECA (International Committee of Museum Educators) Conference. Jerusalem, Israel**:** 10.

Herbjørnsen, O. S. (2003). Software support for creativity, Norwegian University of Science and Technology.

Heymann, P., G. Koutrika, et al. (2008). Can social bookmarking improve web search? Proceedings of the international conference on Web search and web data mining. Palo Alto, California, USA, ACM.

Hildreth, P. and C. Kimble (2002). "The duality of knowledge." Information Research **8**(1): 1.

Hildreth, P., C. Kimble, et al. (2000). "Communities of Practice in the Distributed International Environment." Journal of Knowledge Management **4**(1): 27 - 37.

Chatman, S. (1978). Introduction. Story and Discourse: Narrative Structure in Fiction and Film. , Cornell University Press**:** 19.

CIPHER partners (2001). Enabling Communities of Interest to Promote Heritage of European Regions. CIPHER.

Clark, C. J. (1998). Let Your Online Learning Community Grow: 3 Design Principles for Growing Successful Email Listservs and Online Forums in Educational Settings. Originally Published at the Association for Computers and the Social Sciences (CSS) Annual Meeting. Chicago IL, San Diego State University.

Collao, J. A., L. Diaz-Kommonen, et al. (2003). Soft Ontologies and Similarity Cluster Tools to Facilitate Exploration and Discovery of Cultural Heritage Resources. Proceedings of the 14th International Workshop on Database and Expert Systems Applications, IEEE Computer Society.

Connolly, D., F. v. Harmelen, et al. (2001). DAML+OIL (March 2001) Reference Description, W3C.

Corcho, O., M. Fernández-López, et al. (2003). "Methodologies, tools and languages for building ontologies: where is their meeting point?" Data Knowl. Eng. **46**(1): 41-64.

Cristani, M. and R. Cuel (2004). "Methodologies for the Semantic Web: state-of-the-art of ontology methodology." The Official Bimonthly Newsletter of AIS Special Interest Group on Semantic Web and Information Systems **1**(2): 136.

Crofts, N., M. Doerr, et al. (2005). Definition of the CIDOC Conceptual Reference Model, ICOM/CIDOC**:** 94.

Deerwester, S., S. T. Dumais, et al. (1988). Indexing by Latent Semantic Analysis. CHI'88: Conference on Human Factors in Computing, New York.

Diaz-Kommonen, L. (2002). Design as an activity. Art, Fact & Artifact**:** 150.

Díaz-Kommonen, L. and M. Kaipainen (2002). Designing vector-based ontologies: Can technology empower open interpretation of culture heritage objects? IEEE International workshop on Presenting and Exploring Heritage on the Web (PEH'02) in conjunction with the 13th International Conference and Workshop on Database and Expert Systems Applications, Aix-En-Provence, France.

Díaz-Kommonen, L. and M. Kaipainen (2002). Random Vector Encoding Applied to Cultural Heritage Domain. Helsinki, Media Lab UIAH.

Díaz-Kommonen, L., L. Partanen, et al. (2004). Second trial report - Media Lab/UIAH. CIPHER Deliverables. Helsinki, Media Lab/UIAH.

Douglas, J. Y. and A. Hargadon (2001). "The Pleasures of Immersion and Engagement: Schemas, Scripts, and the Fifth Business." Digital Creativity **12**(3): 153-166.

Edwards, P. (2007). Why Ontologies are Only Part of the Answer for Humanities & Social Science. Proceedings of the Third International Conference on eSocial Science, Manchester University.

Farquhar, A., R. Fikes, et al. (1996). "The Ontolingua Server: a Tool for Collaborative Ontology Construction." Journal of Human-Computer Studies.

Bibliography

Anderson, C. (2004) "The Long Tail " Wired **Volume**, 5 DOI:

Anderson, C. (2006). The Long Tail: How Endless Choice Is Creating Unlimited Demand, Random House Business Books

Arias, E., H. Eden, et al. (2000). "Transcending the Individual Human Mind - Creating Shared Understanding through Collaborative Design." ACM Trans. Comput.-Hum. **7**(1): 84 -- 113.

Aubrecht, P., L. Král, et al. (2004). D18: Collaborative Discovery. CIPHER deliverables. Prague, CTU.

Baxter, H. (2002). "An Introduction to Online Communities." Retrieved 10 June 2004, from http://www.knowledgeboard.com/library/online_communities_introduction.pdf.

Berners-Lee, T. (2001). Webizing existing systems, The World Wide Web Consortium (W3C).

Berners-Lee, T., J. Hendler, et al. (2001). "The Semantic Web." Scientific American.

Brooks, K. M. (1997). Do Story Agents Use Rocking Chairs? Proceedings of the fourth ACM international conference on Multimedia, Boston, Massachusetts, United States, ACM Press New York, NY, USA.

Brown, J. S. and P. Duguid (2002). Learning - in Theory and Practice. The Social Life of Information. Boston, Harvard Business School Press**:** 141.

Brown, J. S. and P. Duguid (2002). Practice makes process. The Social Life of Information. Boston, Harvard Business School Press**:** 106.

Brown, J. S. and P. Duguid (2002). Reading the Background. The Social Life of Information. Boston, Harvard Business School Press**:** 191 - 192.

Bruckman, A. (1996). Finding One's Own in Cyberspace. MIT Technology Review Magazine.

Bush, V. (1945). As we may think. Atlantic Monthly. **176:** 101-108.

Carotenuto, L., W. Etienne, et al. (1999). CommunitySpace: Toward Flexible Support for Voluntary Knowledge Communities. "Changing Places" workshop on workspace models for collaboration, London.

Ceusters, W., B. Smith, et al. (2005). "A Terminological and Ontological Analysis of the NCI Thesaurus." Methods of Information in Medicine **44**: 498-507.

Chandrasekaran, B., J. R. Josephson, et al. (1999). "What Are Ontologies, and Why Do We Need Them?" IEEE Intelligent Systems.

Charles, D. (2001). In The Beginning Was The Word: Making English Heritage Thesauri Available On-line. cultivate-int.

Chatman, S. (1978). Introduction. Story and Discourse: Narrative Structure in Fiction and Film. , Cornell University Press**:** 32.

suggested here with new methods, such as adaptive and personalised technologies, a more flexible, mutable, representational and illustrative approach to web content may be possible.

For example, researchers working on the *Smart Tea* project (see section 6.3.3), when undertaking the knowledge modelling process, used the metaphor of a scientific experiment to bridge the gap between the scientific and computer science communities. A shared understanding is central to a successful ontology, and the ontology designer must work closely with the community when explicitly organising a collective knowledge base. For an effective collaboration strategy to work it should be incorporated into the everyday practices of the community. In this way the community will not see it as an additional burden or overhead. Knowledge is organic and changes as the community learns and re-interprets their domain.

8.3.3. Simple versus Structured ontologies

There is still much research to be conducted before ontologies become a consistent part of the internet. As discussed in chapter 1, a major barrier to the uptake of the semantic web is the cost and effort in developing formally structured, machine-readable ontologies. Outside of this, evaluation is crucial. If an ontology is to be relied upon in a different context from which it was developed, as illustrated by the NCI thesaurus discussed in section 8.3.3, it must undergo rigorous testing and evaluation. Ontologies, as discussed in section 7.3.1, must be developed using accepted theory and practice. In this context there is much scope to study the evaluation and testing of formal ontologies.

8.4. Final note

The research presented in this thesis introduced several approaches to representing collective knowledge. Both the popularity of social software and the proliferation of user-generated content illustrate a need to organise web content according to the user-community. Therefore, by combining many of the approaches terms that gain little currency or are outside the general scope of the domain, tend to go unused.

8.3. Future work

There is much ongoing research into the social web and the area of collective knowledge representation. Indeed the work in this thesis, although only discussing some aspects of a burgeoning research area, does present several possibilities for further research.

8.3.1. Community Interpretation

The approaches discussed in this thesis rely on users contributing explicitly to a collective knowledge base. However, there is much ongoing research into the use of implicit methods to gather user data, though their activity in a forum for example. This can be then be used to produce more semantically accurate domain representations. In this way the community is not burdened with lengthy development cycles but their activity in the forum itself informs their 'world view'. Fields of enquiry, such as collective intelligence, show that there is huge scope to study community activity in this way. Indeed community feedback and implicit analysis is one way that search algorithms are continuously improved.

8.3.2. Community Definition

From the standpoint of the user, there is further scope for understanding the process of collaboration between a community and a designer when developing domain ontologies. Although there are collaboration methodologies available (see section 2.2.3) there is still room for researchers to understand the dynamics of the community model, and create specific strategies and methodologies for those models.

approach is simple, has a low barrier to entry and any user, with a basic understanding of the concepts involved, can immediately start contributing to the overall, collective interpretation of their domain. The same, it could be argued, could be said of the soft-ontology model (see section 5.1). The user is not at any point burdened with technical jargon or lengthy modelling practices. They are, however, participating in the community. It is an active role and their contribution is immediate and considered valid.

In contrast more formal approaches to knowledge representation (see section 2.2.2) involve a more considered and often lengthy modelling process. The community, if non-technical, as was the case with many of the communities considered in this thesis, cannot have as active a role as with other less-formal approaches. Often the information designer must interpret the community's understanding and develop a domain representation, or ontology, accordingly. The approach has the advantage of more acute semantic definitions (as discussed in chapter 7), but the community must see a tangible benefit from the effort. It is difficult to successfully involve a community, however small, in the ontology engineering process; a process that often involves lengthy phases of knowledge acquisition, development and continued maintenance. Implementing the ontology is another problem, as evaluation must be undertaken by the community and further refinements or additions will also involve the information designer or ontology developer. In contrast, a more collective and open approach, such as illustrated by the soft ontology model, develops organically. The community is effectively in control of their domain. They can develop their world view as they see fit. Furthermore, evaluation comes in the form of use: terms that prove to be of little interest, that is retain consistency across a knowledge base, is one simple but effective example. Ontologies, when applied on a grand scale, can encompass thousands of concepts and interrelated properties. Developers, therefore, need assistance as it is extremely difficult for humans to maintain, refine or even comprehend such large datasets. The example of the *Story Fountain* tool, described in section 7.3.2, serves to illustrate one of the principle reasons behind the emergence and popularity of structured ontologies - the explicit statement of that which is implicit.

Naturally, in a field that is new and emerging, there are always further avenues of possibility. It was in this context that interoperability, as discussed in section 7.3.3, was included to illustrate the difference between structured ontologies and simple vocabularies. The former proposes an automatic approach to handling information while the later is considered a useful tool for structuring content, and discourse, for human consumption.

8.2. Structuring Collective Knowledge

There are a variety of ways to approach collective knowledge representation as discussed in this thesis. Many of the approaches, however, will be dictated by the requirements of the knowledge model itself and the ability of the community to contribute to its development. The social web, or ‘web 2.0’, has shown that community members are willing to contribute to the overall organisation of user generated content. Moreover, the rise and popularity of social tagging has been shown by sites such as Delicious. However, currently the resulting taxonomies, often called folksonomies, provide a limited range of functionality. They are, after all, still keywords taxonomies, or metadata subject headings, and do not exhibit strong semantics (see section 2.2). This is, it could be argued, one of their advantages. The indeed contribute successfully when developing their own domain interpretation. Similarly using controlled vocabularies as a strong foundation to develop more specialised taxonomies is an approach successfully demonstrated with the Explorer forum. However, controlled vocabularies are not specifically structured ontologies, as defined by Gruber and Fensel in section 2.2.2. They are subject-based hierarchies without typed instances or property-based relationships. They are flexible, somewhat mutable and can be easily refined or indeed extended. Structured ontologies, unlike tagging frameworks, are not easily extended. They are, for the most part, inflexible and require much consideration when being refactored or re-developed. Therefore, the immediate feedback of approaches such as tagging cannot be employed when using more structured ontologies. To this end it could be argued that the more complex an ontology becomes more difficult it is to directly involve a non-technical user community.

8.1.1. Simple versus Structured ontologies

RQ3: *What impact can the creation and implementation of a structured ontology have when representing narrative concepts?*

There are some clear advantages to developing formal conceptual models such as ontologies, as discussed in chapter 7. However, as highlighted in section 7.3.1, the ontology must be well formed, and not developed without consideration of accepted principles. If the semantic web is to flourish, it is important to develop ontologies with some intellectual rigour, thereby providing a sound foundation for continued and successful maturation.

There are, as discussed in section 7.3.2, some real and tangible benefits to adopting ontologies. Consistency checking, a process where reasoning software can vocabulary or natural language does present the problem of polysemy or synonym bloat: multiple terms being used to describe the one concept, and is more useful when the community is broad and active, otherwise there may not be enough information to describe the domain in a useful manner.

8.1.2. Community definition

RQ2: *What factors can influence the process of knowledge engineering when involving a community of non-technical users?*

Several determining factors, characteristic of user communities (regardless of whether they do or do not meet a physical capacity), were identified as having a bearing on the process of knowledge engineering. Population, for example, will impact the engineering cycle in a number of ways. If the community is excessively large, it is difficult to involve all members during the knowledge engineering process. It can be more manageable to identify a smaller or representable section of the community with a view to creating a model for broader use. Conversely, as discussed in section 6.3.2, the information designer could approach a more focused or expert community, such as a CoP (see section 3.2), with the aim of developing a knowledge model for use in a broader, less focused CoI (see section 3.2). This approach will attempt to take advantage of the stock of knowledge often found in well focused learning communities, such as CoPs. It could be argued, however, that neither of the above approaches is wholly democratic: not every member is afforded the opportunity to participate in modelling the community's domain of knowledge. This, in turn, introduces the broader question of formally specified knowledge versus collective understanding. The rise in popularity of tagging and the corresponding development of the folksonomy, as an alternative to the taxonomy, indicates that communities can approaches to classification provide an effective way for a community to organise their domain of knowledge.

Furthermore the soft-ontology layer and the soft computing approach has the ability to handle and indeed illustrate uncertainty. Traditional methods of classification, as discussed in section 5.3.1, have always presented the user with the problem of deciding in which class a specific artefact belongs. This problem becomes more acute when the user is not offered the opportunity to choose more than one class. Classes are traditionally rigid and do not overlap. Therefore the user must be aware of the fact they are choosing one class over another. Soft-computing, however, is not developed using the same conceptual model. It caters for fuzziness, which, as suggested by Taleb, and discussed in section 5.3.1, is integral to contemporary classification methodologies. Indeed, it is within this context that Clay Shirky maintains that rigid classification schemas, such as the Dewey Decimal System, have become anachronistic when considered within the democratic medium of the web and the ability to include a collective opinion.

In this context, it could be argued that traditional approaches to classification often try to impose a worldview on a group or community. This is because the community do not develop the classification schema but rather adopt a standard approach to organising their domain of knowledge. Indeed, and as discussed in section 5.3.1, this can increase the cognitive overhead experienced by users when classifying artefacts or resources. In contrast, the soft-ontology approach does not attempt to impose a worldview on the community. The method of contribution is simple: natural language statements defining concepts and related properties, and can be employed readily by non-technical communities. The method of contribution presents a low barrier of entry for non-technical members. However using an open

8. Conclusions and Future Work

This chapter presents the conclusions to this thesis. The chapter firstly introduces the contributions to knowledge, in the form of research questions answered. Secondly, the chapter presents a brief summary covering some of the more pertinent aspects of this study. Finally the chapter concludes with suggestions for future work that have emerged from this research.

8.1. Research Questions and Contributions

The background to this thesis, as presented in chapters 1, 2 and 3, raised a number of questions (discussed in chapter 4) that helped to guide this research. Each question, and its related contribution, is now discussed.

8.1.1. Community Interpretation

RQ1: *What are the difficulties of using traditional classification methodologies when approaching community-based platforms?*

The community has become an effective tool when organising and structuring web-based content. Indeed, many of the problems facing the individual in the process of classification can be diminished with the aid of a willing community. Concepts such as The Long Tail (see section 5.3.2) have emerged from studies of collective and social activity on the web. In this context, collective approaches, such as the soft-ontology layer, allow for, and consider evenly, every opinion in a community, and process that opinion according to the collective interpretation of the community. Therefore, a web resource or, in the case of the soft ontology layer, an artefact is classified according to the community's collective opinion. In this way collective the term interoperability discussed in section 7.3.3 serve to illustrate the underlying difference between simple ontologies, as outlined in section 2.2.1 and more expressive, structured ontologies, as discussed in section 2.2.2. The former are developed chiefly for use by human users to organise and retrieve content, while the latter are developed to support software that implement inference services for more sophisticated ontology applications.

Developing a formal ontology introduces the broader question of knowledge, and the representation of knowledge as discussed in section 7.3.1. For an ontology to be useful, it could be argued that the knowledge it embodies should be accurate and developed upon a body of generally accepted theory or practice. In this regard, efforts such as the CIPHER narrative ontology propose a rigorous foundation to develop a more expressive ontology consisting of well-defined semantics. In contrast, less formal or ad hoc approaches to data modelling are developed more rapidly and without the overhead associated with the creation of highly expressive ontologies. Therefore, while structured ontologies can help with consistency checking and interoperability, they must be developed with some rigour and consist of well-formed and generally accepted concepts.

The next chapter presents some conclusions from this thesis. The chapter revisits the research questions, and presents some contributions to knowledge. The chapter concludes with a discussion on future work.

ontology. Both approaches, however, provide users with the ability to semantically browse and search a collection of annotated narratives.

In answering RQ3, and as illustrated by the various implementations of the CIPHER narrative ontology, there are advantages to representing narrative through structured, expressive ontologies as opposed to less formal object orientated data models. Large scale narrative bases can be maintained consistently when represented by a structured ontology and implemented with reasoning software[73]. This can help to eradicate errors when, developing the ontology and later annotating newly created narrative. It is important to maintain highly accurate knowledge bases, especially when working with expressive ontologies. Otherwise errors will arise giving to false assertions. However, as illustrated by the Explorer approach to narrative, this is not possible with simpler ontologies, such as controlled vocabularies or thesauri.

The benefits of interoperability were discussed in section 7.3.3. Although 'equality axioms' are being proposed as a way to align heterogeneous data sources, there is still much work required before systems can interoperate 'intelligently'. In contrast, less formal terminological ontologies, such as the glossary of Irish time periods, can act as a dynamic tool to help multi-disciplinary communities develop a better understanding of a specific domain. Vocabularies, when used in this way, provide a simple yet effective ways to structure discourse. Whilst there is still much research required before systems can interoperate 'intelligently', developing formal ontologies can only help to advance work in this area. The different interpretations of

[73] There are several reasoning engines in use today. Some are freely available such as Jena http://jena.sourceforge.net/ for rule-based inference across RDF or with a licence such as Racer http://www.sts.tu-harburg.de/~r.f.moeller/racer/ for reasoning capabilities across OWL based ontologies.

scientists to appreciate the depth of knowledge required to create well-formed cultural heritage vocabularies, while further offering a deeper insight into the discipline of archaeology. Indeed, from the out-set of the CIPHER project a dynamic glossary was created to act as a foundation for further project discourse. This helped to guide interaction and avoid miscommunication between stakeholders from different disciplines.

Interoperability can occur at several levels and depends on the goal of the implementation. The proposed semantic web is to be founded upon formal ontologies (Maedche and Staab 2001) yet developing formal ontologies is a difficult exercise and the result can often be difficult to implement. In contrast, less expressive vocabularies are a common approach to organising information and while providing a set of domain terms to support discourse, further propose a beginning to interoperability of disparate datasets. This of course is predicated on standardisation and the use of standard vocabularies across a range of data sources.

7.4. Conclusions

This chapter presented two approaches to representing narrative concepts and the domain to which they relate. The first approach was influenced by the earlier efforts at hypertext narrative fiction. Each narrative was divided into smaller textual units called stories, and each story was annotated, through authorial choice, by terms from the underlying domain vocabularies. The approach was implemented as an object orientated data model. The second approach resulted in the CIPHER narrative ontology, a formal ontology developed using the CIDOC CRM data standard. In drawing on traditional theories of narrative, researchers at KMI used several definitions to clearly identify the concepts and properties of the CIPHER narrative

1999), proposes a beginning to ‘intelligent’ or automated interoperability of heterogeneous data. The task itself is difficult and as the semantic web is only in its infancy, and without a bedrock of well formed, formal ontologies and widely utilised tools, there is still much work required. Nevertheless, the notion of automated or ‘intelligent’ interoperability becomes more feasible as interest in the use of structured ontologies increases.

Conversely, vocabularies, or more terminological ontologies as implemented in the Explorer forum, can immediately help to mitigate the problems of communication and interoperability. Controlled vocabularies, for example, can help disparate communities to organise a consensus on a given domain or set of tasks. This is particularly useful when dealing with multi-disciplinary communities, such as ephemeral communities (see section 3.2), who come together for the duration of a project. Although, multi-disciplinary communities can produce new insights, ideas and artefacts, they can also present a possible barrier to progress, particularly at a project’s commencement. This is often because stakeholders, while productive in their own particular spheres are far removed from a collective understanding, a situation Rittel originally described as *symmetry of ignorance*. Arias refines this as the resolution of a problem represented by tacit knowledge in the minds of individual stakeholders (Arias, Eden et al. 2000). Vocabularies, however, help to externalise knowledge in a shared context. This occurs in a balanced and democratic way as the process of developing a controlled vocabulary does not preclude any members of the community, such as non-computer scientists. Indeed, refining the English Heritage thesauri and creating a glossary of time periods for the Explorer forum, helped to structure discourse surrounding the principles of conceptual modelling and the broader and more difficult aspects of ontologies. Furthermore, it helped the computer informal specialisations of broader terms. Therefore, inconsistencies that arise in the Explorer forum's narrative base cannot be identified automatically through the use of reasoning software but instead requires the attention of a human editor. Consistency checking, in this case, serves to illustrate an integral difference between simple ontologies and more complex ones, as the former are developed for human users while the later are structured in such as to be implemented with advanced reasoning software.

7.3.3. Interoperability

A first step towards the semantic web is the task of ontology alignment that negotiates interoperability between heterogeneous data sources, each being represented by formal, expressive domain ontologies. In this regards, McGuinness maintains that '*the detail of expression exhibited by each source's ontology presents the possibility of connecting data sources at a semantic level*' (McGuinness 2002). It is proposed that highly structured, expressive ontologies, such as the Bletchley Park forum domain ontology can make use of equality axioms[72] to define how one concept directly relates to another.

To take a simple example, an ontology may include the definition that a *Bohemian castle* is equal to the concept *castle*, having a typed-instance whose property *place* is filled with the instance *Bohemia*. This definition may be used to develop the concept *Bohemian castle* in an application that does not necessarily understand *Bohemian castles* but does however understand the concept *castle*, the property *place* and the instance *Bohemia*. Although a simple example, ontology alignment (Noy and Musen

[72] Equality axioms stem from logic theory, and donate the equivalence ratio between two sets; i.e. whether set A is equal to set B.

time. To insure, for example, that if event e_1 precedes e_2 and e_2 precedes e_3 then e_1 necessarily precedes e_3.

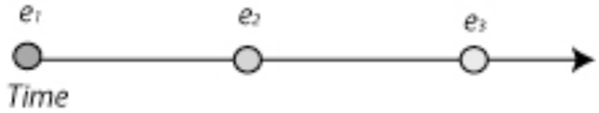


Figure 13: A time line illustrating the relationships between events occuring at different periods. If the event e_1 precedes e_2 and e_2 precedes e_3 then e_1 necessarily precedes $e_{3/.}$

This is an ontological assertion in that it follows the logic principle: Humans are mortal. Greeks are human. Therefore, Greeks are mortal (Shirky 2005). This is a transitive property as discussed in section 2.2.2. The benefit of this sort of reasoning was illustrated during trials on the *Shared Heritage of Central Europe* forum (see section 1.4), when a dozen stories associated with South-Bohemian castles were annotated. The reasoning tool could discover that two stories originating from different castles mentioned a certain person called Oldrich of Rozmberk (Von Rosenberg in German) when in fact these were different people with the same name. The first instance describes how the man died in 1390, while the second describes the man as a confirmed enemy of Hussites, a movement that sprung up after the death of Jan Hus who was burnt in 1415. The inconsistency raised by the two stories could be immediately spotted and alerted to the user by the reasoning tool (Aubrecht, Král et al. 2004).

In contrast, simple or terminological ontologies, as used in the Explorer forum, cannot be used with reasoning software. They do not express sufficient detail to be used in this way. Terminological ontologies are not made up of assertions, but rather contain terms or classes grouped together by similarity. Thesauri hierarchies, for example, are not formal in the sense of strict taxonomies but are rather thought of as consideration. To this end, they suggest, software development is a more pragmatic form of conceptual modelling.

Ontology development is unlike software development in that the principle of the approach is to capture the essence of a concept, as in the example of a story in the CIPHER narrative ontology being represented as '*as a system of associations between elements, composed of events, people and things'*. Researchers working on the Explorer forum did not attempt to capture the essence of a story, but rather develop a system whereby users could explore the domain as an unfolding and dynamic narrative. Moreover, the system was developed for a single instance. In contrast, researchers at KMI proposed a method that allowed a more universal ontology of narrative and a reusable narrative structure.

7.3.2. Consistency checking

Large ontologies may consist of several thousand concepts, all of which contain properties, some of which are confined by a range of value restrictions. This complexity is difficult to model and to maintain consistently. Formal ontologies, however, can help developers and later content managers to maintain consistency across well-structured knowledge bases. This is predicated by the use of a software reasoning system to help identify inconsistencies during development, and later, errors when the knowledge base is in use. The *Story Fountain* tool, for example, on which the *Bletchley Park* forum was based and the CIPHER narrative ontology implemented, was adopted by researchers from the Czech Technical University in Prague to maintain consistency amongst stories developed as part of the *South Central Bohemian* forum (see section 1.4.4). The tool was used to identify inconsistencies in or through practical empiricism. In this way, the knowledge an ontology and its associated knowledge base captures is consistent and comprised of well-founded axioms. The same, it could be argued, is true for glossaries and thesauri; definitions must be correct, synonyms accurate and hierarchies consistent. Building ontologies from scratch is a resource intensive process. Therefore, often, as identified in section 2.2.3, ontologies are assembled from existing ontologies held in libraries such as the DAML Ontology Library[70]. This reduces the overhead of developing the ontology from scratch while also serving to help standardise the domain.

It is within this context that the examples in this chapter help to illustrate the variations in the application of data modelling and the creation and implementation of ontologies. The CIPHER narrative ontology, for example, was founded upon traditional theories of narrative and therefore proposes a step towards the creation of a more universal ontology of narrative. The knowledge and broadly accepted definitions of several theorists (Chatman and Brookes as discussed in section 2.4) helped to create the foundation of the ontology. The ontology could be reused in different instances and, if implemented with a different upper domain ontology (SUMO[71] for example), be applied to different domains. In contrast, the Explorer approach to narrative was developed to function as a specific data model within a single instance, i.e. to associate stories with concepts from the underlying domain. The semantics of a data model, as identified by Spyns et al. and discussed in section 2.2.3, often involve an informal agreement between developers. They argue that amendments occur when warranted and sometimes without large amounts of

[70] The DAML ontology is available at http://www.daml.org/ontologies/

[71] The Suggested Upper Ontology is available at http://www.ontologyportal.org/

perform semantic queries across the narrative-base. In the case of the Explorer forum, the user could explore the domain along three separate axes. Choosing a time period, for example, and narrowing the search criteria with terms from the other two vocabularies, questions emerged such as, '*What tools were used in the Bronze Age?*' The user then explored the domain through a series of unfolding narratives annotated with terms from the underlying vocabularies. The approach was developed in such a way as to accommodate additional vocabularies at a later date.

In contrast, the approach of the researchers at KMI supported the user in asking more specific questions[69], such as '*What was life like for wartime workers?*' or '*How did Bletchley Park influence the development of computers?*' The forum supported additional exploration facilities as described in (Mulholland, Collins et al. 2004).

The following discussion will examine both approaches with respect to research question, RQ3: *What impact can the creation and implementation of a structured ontology have when representing narrative concepts?*

7.3.1. Reusable narrative structures

An ontology should, according to both Gruber and Fensel (see section 2.2.2), embody knowledge that is developed through a shared consensus. However, from the outset, it could be argued that this knowledge should be derived from accepted theory

[69] The ability to ask more specific questions hinges on the level of detail as exhibited by a particular ontology. This brings to light one of the central tenets of domain analysis, in the context of this thesis: identifying the most appropriate, most pragmatic and most useful approach to specifying a domain. The vastness of the Irish domain, for instance, ruled out the use of a highly detailed specification. Conversely, the Bletchley Park forum concentrated on the activities surrounding Station X, the code breaking facility in the Second World War. In this context, it is important to develop an ontology to represent a specific domain or fulfil a specific set of tasks. Over-developing an ontology may burden users when to comes to both annotation and maintenance. It may even require the introduction of specialised data-handlers to preserve the knowledge base. Therefore, as advised by Uschold & King, it is important to clarify the reason, intention or indeed purpose for developing a specific ontology and what are its intended uses (Uschold and King 1995).

Figure 12: A section of the CIPHER Narrative Ontology. Researchers at KMI extended the CRM to develop a more formal definition of narrative and its related concepts. The CRM classes are coloured green, and the newly created narrative ontology classes are coloured yellow.

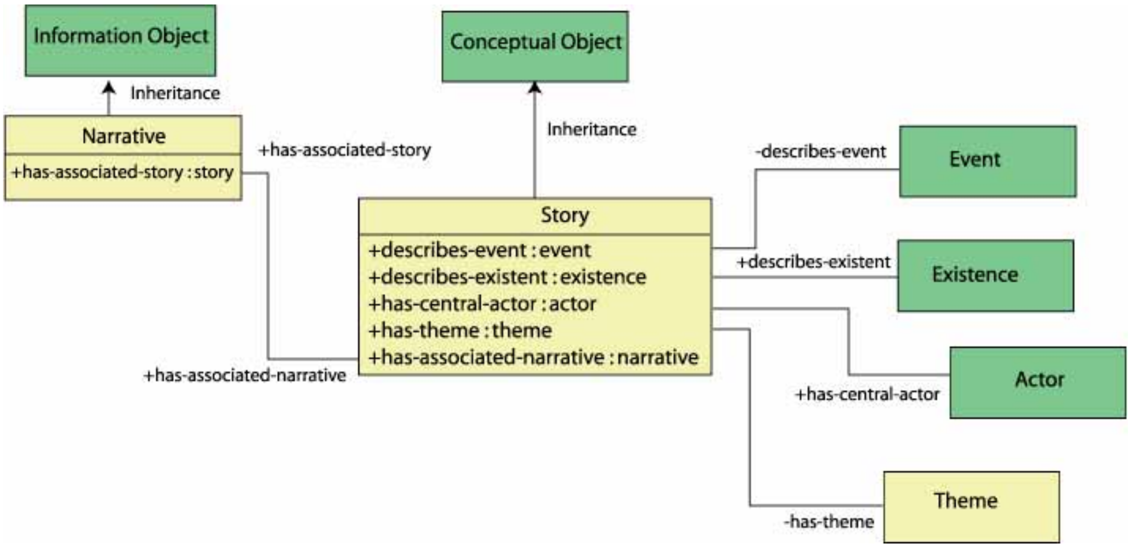
Figure 12 illustrates the approach taken by researchers at KMI in using the CRM as a foundation to develop a more formal definition of narrative. They extended two CRM classes; E73-Information Object[68], to define the concept of narrative, and E28-Conceptual Object, to define that of story. They further extended the CRM to include the concept discourse. As the notion of time (and of events marking specific points in time) is an intrinsic aspect of the heritage domain, the CIDOC team took an event-centric approach to modelling the ontology. Researchers at KMI choose the CRM event class, E5-Event, to represent the concept of an event in the CIPHER narrative ontology. Similarly, they adopted the CRM class, E39-Actor, to represent the concept of an actor (in the sense of a person), and E77-Existence to represent the concept of existent in the CIPHER narrative ontology. The result was a conceptualisation of narrative reflecting traditional narrative theory and developed using the CRM upper domain ontology.

7.3. Formally Structured Knowledge

While the first approach, implemented in the Explorer forum (see section 1.4.1), sought to present the underlying domain through a series of dynamic narrative presentations, the second, based on the Story fountain tool (see section 1.4.3), aimed to represent narrative concepts according to traditional narrative theory. Both approaches aimed to explicitly represent the narrative form, supporting users to

[68] All CRM classes are described in full in (Crofts, Doerr et al. 2005)

This CIPHER narrative ontology was implemented as an extension of the CIDOC Conceptual Reference Model (CRM). The CRM[65] is the culmination of over twelve years of development by the members of the CIDOC team and represents a formal ontology to facilitate the integration, conciliation and interchange of heterogeneous cultural heritage information (Crofts, Doerr et al. 2005). The CRM is an object-orientated semantic model with formal subclass hierarchy and typed instance creation, and could therefore be found towards the right hand side of McGuinness' spectrum[66] (McGuinness 2002) (see also Figure 4). The CRM does not aim to define the terminology of a respective domain but rather suggest the relationships for its use. It could therefore be classed as an upper domain ontology[67] describing the overarching concepts of a broader domain, in this case the domain of cultural heritage.



[65] For more information on the CIDOC CRM see http://cidoc.ics.forth.gr/

[66] The CRM is an object orientated, property-based model with both symmetric and inverse properties as defined in the original specification (Crofts, Doerr et al. 2005).

[67] An upper domain ontology can be classed as an ontology describing more general concepts relevant to s specific domain. In this way the ontology can act as a foundation, in an ontology stack, for more specific domain ontologies. The CRM in this case does not try to describe the domain but rather provide a conceptual structure for inter-relation of more specific cultural heritage domain ontologies.

Researchers at KMI used traditional theories of narrative to help structure and define the CIPHER narrative ontology (Mulholland and Collins 2002; Mulholland and Zdrahal 2003; Mulholland, Collins et al. 2004). In reflecting both Brooks' and Chatman's definition of a narrative as the '*manner of expression*' or the way in which a story is told (see section 2.4), they created an ontological class to represent the concept of narrative. Similarly, reflecting Brook's definition of a story as '*as a system of associations between elements, composed of events, people and things*' (see section 2.4), they created a second class representing the concept of a story.

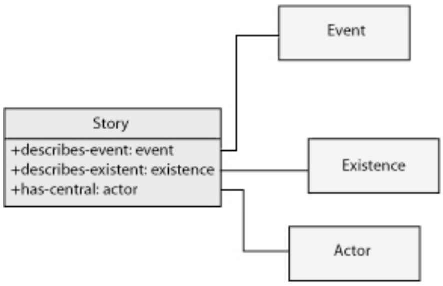


Figure 11: The concepts Story, Event, Existence and Actor as reflected in the Cipher Narrative ontology.

Figure 11 illustrates the skeleton of the CIPHER narrative ontology. Delving further into narrative theory, however, Mulholland et al. began to sketch out a more thorough semantic definition of what constitutes the narrative form. For example (as discussed in section 2.4), Chatman proposes that narrative is not a disconnected amalgam of events and existents but rather manifests itself in a coherent and structured manner. Additionally, Chatman suggests that a story has an author, is told by a particular narrative, and is comprised of story elements, or events and existents (objects, individuals, settings, characters) (see section 2.4). Furthermore, discourse was defined as the way in which a story is told. Discourse, in this sense, determines how elements of a story are presented in terms of style, ordering of story elements and media used to emphasise the narrative.

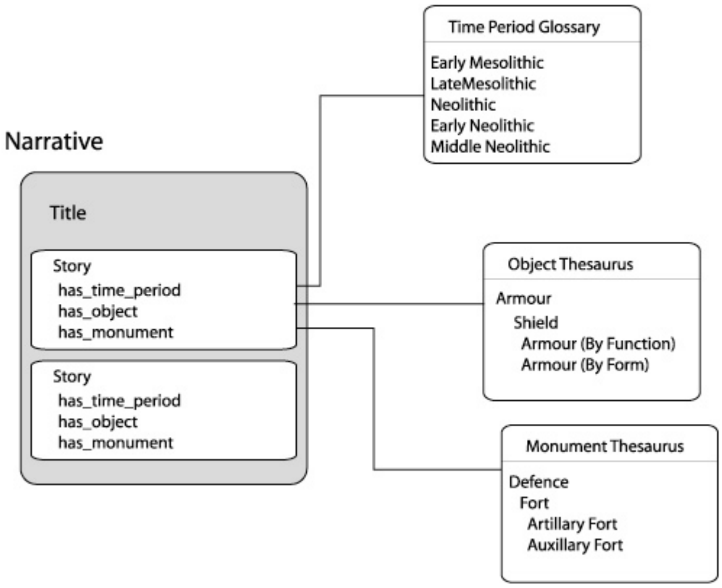


Figure 10: A representation of the Explorer approach to Narrative. Narrative contained a number of stories each annotated with several terms from the thesauri/glossary domain representation.

Figure 10 illustrates the Explorer approach to narrative. Narrative is divided into discrete story objects, each with relationships to the glossary of time periods, the Monuments Types Thesaurus and Archaeological Objects Thesaurus. The author was provided with the ability to make associations between each story and the underlying domain. They could further associate stories in a linear fashion to produce a complete narrative presentation.

7.2. The CIPHER Narrative Ontology

The CIPHER narrative ontology was implemented in the Story Fountain, a software tool developed to support the Bletchley Park tour guides organise narrative content (see section 1.4.3). The Story Fountain tool was also implemented in the *Shared Heritage of Central Europe* (see section 1.4.3) by researchers at Czech Technical University, Prague. The following section will discuss the approach of researchers at KMI when developing the CIPHER Narrative ontology.

standpoint a story is a piece of text, which relates to some aspect of the domain (Kilfeather, McAuley et al. 2003). It can be as small as a few lines discussing a single domain instance, e.g. Trim castle (Trim castle in this example is a domain instance of the class Castle from the Monument Type Thesaurus). The principle involved dividing narrative into smaller discrete units that were identified as stories (Figure 10), each representing specific domain instances from the underlying knowledge base[64], or, indeed, terms from underlying vocabularies. The reader can investigate, explore, and navigate the domain via the annotations and ontological relationships within the narrative layer. This approach somewhat reflects that of the hypertext fiction (see section 2.4.2) in that the reader is immersed into the overarching narrative, as each step into the narrative base produces a range of similarly themed stories for the reader to explore. It is, however, a more formal approach than that of hypertext fiction (see section 2.4.2), as unlike the open link structure of a hypertext narrative, the link structure of an Explorer narrative is defined by terms and relationships from the underlying vocabularies. The reader is therefore exploring a predefined and closed domain.

[64] Knowledge base in this context refers to a well structured database, organised in such a fashion as to support the retrieval of knowledge or, in the case of the Explorer forum, annotated narrative and media content.

7. Simple versus Structured Ontologies

This chapter investigates the use of simple ontologies, as outlined in section 2.2.1, in contrast with more formal, expressive ontologies, as described in section 2.2.2. Two approaches, the Explorer approach to narrative and the CIPHER Narrative ontology, are used to help examine some of the tangible benefits to creating and implementing structured ontologies. The discussion will, however, initially explore the differences between representing narrative for a single instance compared to creating reusable narrative structures. The reasoning behind this is to outline the disparity between more ad hoc approaches to data modelling, as are often undertaken by a software developer, and formally structured knowledge, as created and represented by domain experts and knowledge engineers. The chapter will go on to discuss each approach under the headings: reusable narrative structures, consistency checking, and interoperability. This chapter concludes with a summary and brief discussion on the benefits and limitations of approaching narrative through the use of formal ontologies. This chapter will therefore consider the following research question, RQ3: *What impact can the creation and implementation of a structured ontology have when representing narrative concepts?*

7.1. Explorer approach to narrative

Explorer (see section 1.4.1 and Appendix B: Developing Explorer) was developed with the aim of firstly, allowing members of a CoI to contribute narrative and media content to the forum, and secondly, providing a mechanism to support the community to explore Irish heritage as a series of dynamic narratives. To this end, and partially reflecting Landow's work on hypertext narrative fiction (see section 2.4.2), researchers at DIT introduced a basic narrative unit called a story. From a user's and one that involves a large investment from the community of users. Introducing the concept of formal ontologies to non-technical communities (communities whose focus lies within the humanities or social sciences for example) will not always produce a successful outcome. Rather, the community should be involved at every level of the engineering process. In this way the representation belongs to the community and evolves with the community's understanding. Such a representation is not rigid and static but rather mutable and dynamic, and can therefore evolve as the community grows and develops.

The next chapter will explore the application of structured ontologies. Although, there is an amassing amount of literature on the development, maintenance and use of ontologies, the discussion will focus on two specific real examples. While the first example involves the creation and application of a standard vocabulary, the second involves the creation of a structured ontology. The chapter will use both examples to highlight some of the disparities between the development of ontologies and the creation of simpler data models, and to investigate the advantages of adopting a structured ontological approach to representing the narrative form.

familiar to scientists[63] (Frey, Hughes et al. 2003.). In both cases, the community of users felt comfortable working with the system and understood the approach to knowledge representation.

6.4. Conclusions

This chapter introduced two different ways to represent collective knowledge. The first resulted in the creation of a glossary, implemented as part of the Explorer forum, which attempted to broadly chart the Irish archaeological timeline. The second involved the creation of a formal ontology, which aimed to represent the activities of Station X at Bletchley Park during the Second World War. Each approach was then used to discuss several aspects of collaborative knowledge engineering, namely: population, community models and collective knowledge versus formal ontologies.

RQ2 questioned whether the population of a community can impact the approach to knowledge representation. If the community is large it is difficult to approach the process of knowledge acquisition and later knowledge modelling in a democratic manner. Secondly, the type of community, be it a CoP or CoI, can further impact the approach to knowledge representation. As CoPs comprise of people who work continuously in a specific domain, they can often be more focused on the completion of a task than CoIs. Finally, knowledge engineering is an epistemological exercise

[63] It could be argued that Frey's paper, titled 'Less is More', follows Richard Gabriel's controversial 'Worse is Better' essay on coding practices. In it Gabriel described a design philosophy which espouses simplicity and piecemeal growth over complexity. Although he received criticism his paper sparked wide debate on design philosophies, it is, it could be argued, applicable to the growth of knowledge languages or ontologies. The premise of his essay was that people will take up simple languages and approaches above complex ones, and can then, in turn, try to develop and improve on these simple approaches (Shirky 2008). Indeed Frey's paper suggests a similar methodology.

researchers at KMI implemented these features through ontological concepts and their interrelationships. Currently, many of the concepts and tools that support the creation and maintenance of formal ontologies require a high level of expertise, and are therefore usually undertaken by researchers with experience in the field. Formal ontologies by their very nature are inflexible and, in a similar way to large scale computer systems, require specialised data handlers or engineers to administer changes or perform maintenance. In contrast, less formal models, such as the glossary of time periods in the Explorer forum, can be created and administered by non-technical communities. The concepts are easy to understand and the benefits often tangible. Indeed the glossary of the Irish Archaeological timeline was thought of as a dynamic tool that users from the broader community could update or specialise through the addition of specific events, such the Battle of the Boyne or the 1916 Easter rising. This approach reflects *Methontology*, as instanced in section 2.2.3, in that the community begins with a controlled vocabulary and then progresses with greater specialisation and more acute semantic definitions. Moreover, the representation belongs to the community and thus evolves with the community's understanding. It is more difficult to approach formal ontologies in this way. Indeed, researchers working on projects such as *Policy Grid* and *Smart Tea* have recognised the benefits of community participation during all stages of knowledge engineering. *Policy Grid*, for example, reorganised an entire system to incorporate social tagging after unsuccessfully introducing several ontologies to a community of social scientists (Edwards 2007). Similarly, researchers working on the *Smart Tea* project developed a lightweight ontology using the metaphor of the experimentation process, a process representation of the Irish archaeological timeline. Similarly, when specifying the domain of Station X during World War Two, researchers from KMI collaborated closely with the Bletchley Park tour guides.

6.3.3. Collective Knowledge versus Structured Ontologies

Outside of the scientific, or indeed technological, disciplines, knowledge engineering and the creation of formal ontologies may not generate comparable interest. The Discovery Programme staff, for example, expressed little interest in using the CIDOC CRM data standard. The CRM was put forward by researchers at DIT during a workshop with the Discovery Programme, and although they understood the advantage of standards and the reasons behind the creation of upper domain ontologies, they could not see the tangible benefits of using such a highly complex model. It could be argued that the CRM is not supposed to be used by domain experts in this way, but rather implemented by engineers as an upper domain ontology and later populated through use in a knowledge-based system. Conversely, the argument could be made that engineers have little experience of the CRM and are therefore not in an advantageous position to implement such a complex model. It is difficult to bridge the disciplines in this way. As the engineer may have little experience of the heritage domain while similarly the heritage professional has little experience of data modelling or the implementation of formal ontologies. Consequently the workshop served to illustrate the larger issue of the creation, implementation and maintenance of formal ontologies by non-technical communities.

Having said this researchers at KMI, when developing the Bletchley Park forum, created a formal domain ontology founded on the CIDOC CRM. In this process, the Bletchley Park tour guides highlighted the most salient features of the domain and

"Cruchain Ai" heritage centre. The argument could be made, however, that in strict accordance to Gruber's definition of ontology (see section 2.2.2), such a two-tiered approach does not include the entire community when representing their domain of knowledge. The approach rather reflects a top-down design methodology in that a small group of experts develop a representation for use in a broader community[62].

6.3.2. Community Models

According to Welty, *quality knowledge-based systems depend on quality knowledge* (Welty 1999). CoPs, for example, tend to have a more thorough understanding of a subject area, as they often consist of practitioners who work continuously within a specific domain (see section 3.2). It is within this context that the interpretation of a group of domain experts, such as professionals working at the Discovery Programme, can provide a solid foundation from which to explore Irish history, while similarly a conversation with a *Bletchley Park* tour guide may transform a casual museum visit into an enlightening experience. On the other hand, CoIs (see section 3.2) tend to have a looser focus as they predominantly consist of practitioners from different domains who are united by a common interest. It was in light of this that researchers from both DIT and KMI approached CoPs when initiating the process of knowledge engineering. Both communities, they felt, were focused and revealed a broad and often thorough understanding of their domain. Researchers at DIT drew upon the knowledge and experience of the Discovery Programme when refining the English Heritage thesauri and creating a valid

[62] A participatory design methodology, as described in (Díaz-Kommonen, Partanen et al. 2004) can have a further impact. Díaz-Kommonen et al. suggest that incorporating a CoP, or indeed several CoPs, can help develop a strong foundation for the emergence of a wider CoI. This, the author suggests, can include creating specific domain representations and seeding the forum with relevant and interesting content.

6.3.1. Population

In (Shirky 2008), Clay Shirky describes the *birthday paradox* as a means to illustrate the difficulty of coordinating communities. The birthday paradox suggests that a group's complexity grows much faster than its size. Therefore, it is unfeasible for everyone to interact directly and more difficult to organise the group. In this context it could be argued that the population of a community can directly impact the approach to knowledge engineering. This, of course, is dependent on the information designer incorporating the community into the engineering process. If the community is excessively large, as in the case of the visitors to "Cruchain Ai" heritage centre in Tulsk, it is difficult to involve all members of the community during the engineering process. In contrast, approaching smaller communities, such as the Discovery Programme, makes this process more manageable and therefore more productive. The community can attend physical meetings and if, for instance, members of the community are unfamiliar with the principles of knowledge engineering (as in the case of the Bletchley Park tour guides) workshops can be held to help introduce the more complicated aspects of the process. Moreover all members of the community can be explicitly involved, each voicing their opinion and therefore contributing to the overall method. In this respect the approach of researchers at KMI proved successful, although the resulting domain representation was never intended for use within a broader community model[61]. In contrast, the approach of researchers at DIT involved collaborating with a focused community, such as the Discovery Programme staff, to develop a set of vocabularies for use in a wider community, such as visitors to

[61] In this instance members of the Bletchley Park tour guides used the domain representation to annotate newly written narratives. People interested in the park could also use this representation to query the narrative base, retrieving stories of interest. In this way the model was restricted to authors who were writing narrative for the system.

continuously on related activities in the Bletchley Park museum. The approach of researchers at KMI reflected the earlier work of Srinivasan (see section 2.2.3). This process involved defining the principal and concrete concepts of the domain, which were insistently recognisable by the participating community. The process began with tour guides listing their areas of interest and using these to map the domain and organise selected content. This activity helps to focus the community on the task in hand while further illustrating, in a tangible way, the process of domain representation. It can be used to bridge each community's understanding, in that the tour guides can immediately partake in and therefore experience the process of domain representation while the researchers can use the activity as a foundation to structure the domain.

6.3. Representing collective knowledge

In 2.2 Gruber defined 'an Ontology as a formal, explicit specification of a shared conceptualisation'. His definition highlights the point that the knowledge an ontology embodies is intended to represent a shared understanding. It is therefore necessary for an ontology to comprise of concepts that are agreed upon by a community. In this way an ontology is not developed by an individual but rather cultivated, specified and later maintained by a dedicated community. However, taking the previous sections as an example of two different approaches to knowledge representation it is important to examine them in relation to RQ2: *What factors can influence the process of knowledge engineering when involving a community of non-technical users?*

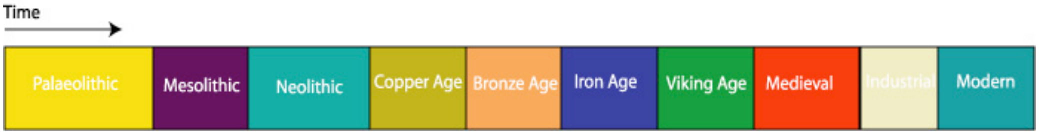


Figure 9: The principle time periods belonging to the Irish archaeological time-line. Several time periods, such as Neolithic, are further divided into early, middle and late. The time line was thought of as a dynamic tool that allowed users to specify more specific periods as the famine in 1847 – 1848 or events such the Battle of the Boyne in 11th July 1690.

There still remained the matter of representing the Irish archaeological record, which begins around 10,000 B.C., and continues through the Mesolithic and Neolithic eras, the Bronze and Iron Ages, culminating with modern times. Members of the Discovery Programme suggested attributing temporal relationships at a period level. The resulting controlled vocabulary contained 27 time periods, such as Early Neolithic, Neolithic and Late Neolithic, and was developed by several members of the Discovery Programme (Figure 9). It was later extended into a glossary with natural language descriptions appended to each time period. However, for such a broad representation of time, it was felt by researchers at DIT that users of the forum should be able to extend the glossary to include more specific, and sometimes seminal, events in Irish History, such as the 1916 Easter Rising. In this way the time period glossary was thought of as a dynamic tool to be continuously refined by users of the forum.

6.2. Collaboration on the Bletchley Park Forum

The principle community involved in developing the Bletchley Park forum consisted of a small group of about 35 volunteer tour guides, who on a daily basis conducted tours of the Station X museum at Bletchley Park. The community is similar to the earlier definitions of a CoP (3.2), in that they work collectively and

Researchers at DIT approached members of the Discovery Programme to discuss the process of domain representation. This decision was taken because the Discovery Programme has extensive experience in the domain of Irish archaeology and in this way constitutes a community of domain experts or CoP. As there were, however, no standard vocabularies or indeed ontologies representing the domain of Irish history, the problem of whether to design a new one or to adapt an existing vocabulary presented itself. During preliminary discussions with members from the Discovery Programme, researchers from DIT introduced the process of knowledge representation and presented several approaches to structuring real world knowledge, such as controlled vocabularies, taxonomies and ontologies (2.2). During these discussions members of the Discovery Programme Staff indicated that they felt most comfortable with more terminological-based ontologies, such as thesauri, because similar approaches are commonly used, for subject-based classification mostly, in the field of archaeology. The breadth of the domain was also discussed, and members from the Discovery Programme Staff suggested representing the domain of Irish history along three distinct axes, time, artefacts and monuments. In this way, a reader could ask questions such as: what farming tools were commonly used during the Bronze Age? The query could return stories annotated with the concepts Bronze Age and Farming tools. As English Heritage, had at the time of development, created several thesauri during a major computerisation project (Charles 2001), members of the Discovery Programme Staff suggested contacting English Heritage and refining the Archaeological objects and Monuments thesauri to suit the Irish domain.

6.1. Collaboration on the Explorer Forum

Researchers from DIT developed the Explorer forum (see section 1.4.1 and Appendix B: Developing Explorer) in collaboration with the members of the Discovery Programme staff. The forum aimed to encourage the public to contribute stories, in the form of narrative presentations, based on aspects of Irish cultural heritage. The approach involved developing a representation of the domain of Irish history and pre-history to support users of the forum to annotate their stories with terms from the underlying vocabularies.

Several communities participated in the creation and ongoing development of the Explorer forum. This thesis, however, will focus on two very specific groups. This is because they reflect the community models of CoP and CoI, as discussed in section 3.2, and can therefore help to illustrate the process of collaboration with two different community models. The first community, the Discovery Programme staff, consists of a group of heritage professionals undertaking research in Irish archaeology and regional heritage. The community type reflects the notion of a CoP, as it is comprised of practitioners who work continuously in the domain of Irish archaeology and history. The second, however, were visitors to "Cruchain Ai" heritage centre[60] in Tulsk, Co. Roscommon. The centre is run as a not-for-profit heritage facility and assists pupils and teachers in exploring the many aspects of Celtic heritage in imaginative and insightful ways. This community, matching the notion of a CoI, was quite large (~300 people) and diverse though it chiefly comprised of student classes visiting the centre.

[60] Information about the Crucain Ai heritage centre can be found at the following URL: http://www.cruachanai.com/.

6. Community definition

This chapter investigates the process of knowledge engineering when applied to non-technical user communities. There are, as evinced in chapter 2, several approaches, such as glossaries, thesauri or formal ontologies, to structuring and representing real-world knowledge. In addition, there are several factors that can impact on whether one approach to knowledge representation is more successful than another. This author suggests that one such factor is the demography of the user community, while a second is the method of collaboration. It is within this context that this chapter introduces two approaches to representing knowledge. The first, developed by members of the Discovery Programme, attempted to chart the Irish archaeological timeline in a concise glossary. The glossary was later implemented as part of the Explorer forum (see section 1.4.1). The second involved a close collaboration between the Bletchley Park tour guides and researchers at KMI in the production of a formal ontology, which aimed to reflect the period surrounding the Second World War and the activities of Station X at Bletchley Park (see section 1.4.3). Each will be considered in terms of their approach to community population, community model and also from the perspective of a collective versus a formal approach to knowledge representation. The chapter concludes with a summary of some of the aspects that impact the process of collaborative knowledge engineering. The chapter will attempt to answer the following research question, RQ2: *What factors can influence the process of knowledge engineering when involving a community of non-technical users?*

in technical jargon and heavy conceptual modelling, yet provides a more harmonious approach to classification. Rigid classification schemas, as instanced by the application of the English Heritage thesaurus, however, impose a specific world view on the community. Such an approach can impede individuals when classifying objects by ensuring that the individual chooses one class over another. Nevertheless, if the domain is limited and the vocabulary fixed, the user is unable to submit further synonyms to the model, therefore reducing the number of terms identifying a single concept or property.

The next chapter will investigate and answer research question RQ2. The premise of the chapter is to discuss how the community, regarding type, form, demography, population and other factors impact the approach to, and outcome of, knowledge representation.

does the outcome represent the interpretation of the entire group? The soft-ontology tool was not necessarily deployed in this way, but it does suggest that representing collective knowledge through the use of social tools may only, in fact, illustrate the interpretation of the most active 20% or 30%.

5.4. Conclusions

This chapter considered two methods of representing a domain. The first, soft–ontologies, was an approach implemented as part of the Carta Marina forum (section 1.4.2) while the second provided support for the classification of content in the Explorer forum (see section 1.4.1). The soft-ontology approach was developed using advancements in the field of soft computing and benefited from techniques such as the SOM algorithm, as discussed in section 2.3.1. In contrast, the Explorer thesauri were adopted from English Heritage and later refined by a domain expert. The Explorer approach, however, reflected more traditional classification schemas and therefore provided a means to contrast the application of soft-ontologies. The chapter compared and contrasted both approaches under the headings: collective classification, method of contribution, community models and limitations.

In answering RQ1 it was recognised that adopting the soft-ontology approach diminishes the need for a top down design methodology (as illustrated by web directories such as Yahoo and Google), which in turn lessens the need to classify digital artefacts according to a strict and rigid classification schema. In this way the information designer is not imposing a specific world view on the community. The method of contribution - natural language statements describing concept properties - reduces the barrier of entry for a non-technical community. This is of particularly significance in the context of this thesis, as the SOL avoids burdening the community

5.3.3. Limitations

A clear advantage in using controlled vocabularies (or more structured metadata models) is that classification is based on a limited set of terms. More open approaches, such as the soft-ontology model, can result in several synonyms being used to identify the same property or concept. For example, one individual may describe a bird as having a beak while another may suggest that a beak is actually a bill leading to polysemy. Spell mistakes can further bloat the sample with redundant terms. Although a spelling checker could be employed to help users from committing egregious spelling errors, there is no clear way to stop people describing artefacts with synonyms, outside of closing or limiting the vocabulary. There are, however, other approaches, such as auto-suggest, which can be used to prompt the user with similar terms while still providing the user with the freedom to describe an artefact as they wish.

The distributions of many social systems, as identified by Chris Anderson amongst others, often follow Pareto's 80/20 power law. This is because power law distributions tend to describe systems of interacting elements. Therefore a power law, and the resulting long tail, can be used to illustrate most social tag distributions, and the author suggests, if applied in a broad ecological instance, the soft ontology model. However, the long tail can also reflect the most active members in a community. Generally, the most active contributor tends to be much more active than the median participant (Shirky 2008). In fact, in a perfect power law distribution, the most active participant is twice as active as the second most who again is twice as active as the third most and so on. Unlike a Gaussian distribution the median is therefore inconsequential as 80% of contribution comes from 20% of the community. If the majority of contributions are coming from a handful of active community members

8. If enough people tag a sword as a shield, that sword will be represented as such in the overall distribution. However, the effect is equally true in reverse. If only a small number of people tag a sword (incorrectly) as a shield it will be so insignificant as to not impact the greater sample.

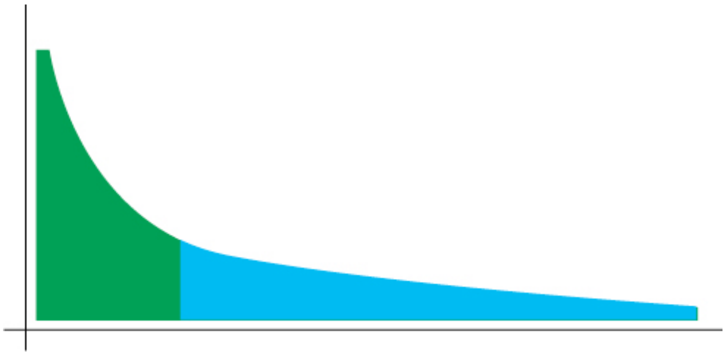


Figure 8: Illustrates the Long tail Graph popularised by Chris Anderson. Clay Shirky maintains distributions that illustrate social or collective interaction adhere to the 80/20-power law or what has become known as 'the long tail'.

The same, it could be argued, is true of a soft-ontology layer. If there are a small number of users in a large community who incorrectly define a sword as soft, their opinions will be drowned out by the collective opinion of the group. However, as mentioned previously, soft-ontologies have an advantage over social tagging in that the final representation is not only collective but also developed on more acute semantic definitions. This is to say that soft ontologies describe the properties of an artefact. In this way the natural language statements used in the soft ontology model are more qualified than tags.

the niche products that people are unable to buy elsewhere. Although, there is much anecdotal evidence surrounding the long tail there have been some quantitative analyses carried out investigating the complex dynamics of tagging and social behaviour e.g. (Halpin, Robu et al. 2007).

content. However, using more detailed or structured knowledge models require more effort from the user, particularly, when annotating new resources.

5.3.2. Community Models

The nature of an online community, it could be argued, can impact the approach to representing the community's domain of knowledge. There are a variety of community models (as introduced in section 3.2), some task based, such as CoPs, others more diverse, such as CoIs. An approach can prove more successful when the community model is taken into consideration. Not all communities are interested in creating or deploying sophisticated ontologies. They may have a looser focus, as exhibited by some CoIs, or comprise of users without a high level of technical expertise as illustrated by Srinivasan's community of refugees and his resulting approach to fluid ontologies (2.2.3). Similar to Srinivasan's Village Voice project, the soft-ontology model presents a low barrier to entry for non-technical communities.

The popularity of social tagging has indicated that less complex methodologies can ensure a greater and more collective involvement[58] when organising community-based content. There are further benefits to broadening the user base in this way. More collective approaches to classification take advantage of the power law distribution recently described as the long-tail[59] (Taleb 2007), as illustrated in Figure

[58] James Surowiecki, in his book the Wisdom of Crowds (Surowiecki 2004), argues that if a group or crowd exhibits enough diversity and independence, then they can act in a more intelligent fashion. He presents much anecdotal evidence in which crowds, comprising of independent and free thinking individuals organised in a decentralised manner, effectively enhance the decision making process.

[59] The long-tail has become a popular phrase to describe the 80/20 power rule, sometimes referred to as Pareto's rule, whereby 80% of the effects come from 20% of the causes. Chris Anderson, in his article and subsequent book *The Long Tail* (Anderson 2004) (Anderson 2006), discusses the impact of choice when demand is unlimited. His thesis is based on the Amazon model whereby 80% of the customers purchase 20% (or the long tail) of the catalogue. This 20%, however, lies outside the most popular books that the average bookseller is likely to stock. Therefore, he suggests, that Amazon profits from

Again, it could be argued that the prevalence of social tagging is predicated not only on its effectiveness but also on the ease in which a member of a community can tag an online resource[56]. Similarly, the soft-ontology approach does not require a member of the community to complete a lengthy annotation phase when describing an object. Annotation takes the form of natural language descriptions of artefact properties and, if required, the insertion of a property weighting, ranging from 0 to 1. Yet, the user is incorporated into the process.

Similarly, researchers working on the Explorer forum developed tools to allow users add content with as much ease as possible. This was to avoid placing barriers between the community and the contribution process. In this context the user could, if they required, classify or indeed annotate their content with as many terms from the thesauri as they desired. It is arguable whether this is a problem for some users. There may be some cognitive overhead[57] in choosing the correct terms or classes to identify a specific resource (as discussed in section 5.3.1). However, the user is free, when using the soft-ontology model, to describe a property as they wish.

While both approaches support broad community models, each introduces the wider problem of poor classification. In attempting to maintain a well-structured knowledge-base, it may be required, in some instances, to employ the services of a content manager to vet contribution and sustain more organised and well-annotated

[56] In (Halpin, Robu et al. 2007) Halpin et al., when discussing the complex dynamics of social tagging, cite Zipf's law, which states that information seekers will use the most convenient method of searching to retrieve an information resource. Users therefore use the tools they are most comfortable with and that are the least challenging.

[57] Rashmi Sinha in (Sinha 2005) suggests the lower cognitive costs of tagging, or the use of natural language statements to describe resources, are one of the reasons for its immense popularity. She suggests that there is a decision making process when using typical or more traditional category based classification schemas (such as a thesaurus). This is further discussed in (Heymann, Koutrika et al. 2008) where the user must be aware of a specific vocabulary to describe or annotate a resource.

methodologies such as the Dewey Decimal system, which Clay Shirky identifies as anarchic when considering the democratic medium of the web and the community's collective opinion (Shirky 2005). Indeed collective approaches to organising or indexing content have become one of the salient features of the 'web 2.0' paradigm. Social tagging, a prominent way for communities to identify interesting or popular content, is common across content driven, community-based websites. Tagging is similar to the soft-ontology model in that users can annotate resources with natural language statements and are not confined by a prior classification schema. Indeed it could be suggested, the user's opinion has never been as valued as it has been by the rise of 'web 2.0'. However, as users are asked to describe artefact properties, the soft-ontology approach has the benefit of presenting a more thorough domain representation than that of social tagging. This is because soft-ontologies support the user to describe and weight properties belonging to a specific object. This bottom up design methodology that was once shelved for fixed taxonomic structures is increasing in popularity, as instanced by the rise of sites such as Blinklist[55] or Delicious.

5.3.1. Method of contribution

A further aspect of this approach is the method of contribution. The importance in the way in which contribution is undertaken was identified in section 3.3. Wikipedia, for example, do not burden the user with a lengthy process of contribution. It is simple, effective, and does not present a barrier to involvement in the community.

[55] Blinklist (www.blinklist.com) takes a similar approach to the other web 2.0 sites such as stumble upon (www.stumbleupon.com) and delicious (del.icio.us) in that it supports a collective approach to categorising content. For a more comprehensive study on the nature and state of the art of social tagging see (Voss 2007).

remove the need to organise objects according to rigid classes. In this sense the approach supports the community in effectively controlling their domain of knowledge. They are not forced to shoehorn objects into one class or another but rather are empowered to develop their domain model as they see fit. This ability helps to diminish the inherent problem of classification of a fixed domain. Nevertheless, it is worth noting that the soft-ontology approach is a formal, yet relatively user friendly, method of creating a properly encoded, machine readable, representation.

From this perspective, Díaz-Kommonen et al. maintain that it is not the task of the information designer to 'chew the world for the user' (Díaz-Kommonen and Kaipainen 2002). The designer, they argue, should develop tools that empower the community rather than restrict it. This was identified as '*open interpretation approach to information design*' (Díaz-Kommonen and Kaipainen 2002) allowing community member's freedom to structure and interpret information according their personal needs. In this way users are not confined by fixed taxonomic structure as illustrated by the more traditional approach of the Explorer forum. As discussed in 5.2, the English Heritage thesauri were firstly developed and later refined by groups of experts with extensive experience in the domain of cultural heritage. It was proposed that each thesauri should remain as close to the original standard as possible. In this way the broader community could not suggest amendments or additions to either vocabulary. In effect, this approach reflects a top-down design methodology in that a small group of domain experts develop thesauri to be used by a broader community of users. Taleb proposes that classification is necessary in this context but can become problematic when thought of definitive, preventing people from considering the fuzziness of boundaries (Taleb 2007). Similarly, thesauri reflect

5.3.1. Collective interpretation

There is an inherent problem with traditional approaches to classification[54]. This problem can be illustrated by asking two people to classify objects by their hardness. Often, in such situations, traditional classification schemas may fail to express the collective opinion of both involved. One person, for example, may think the object is soft while the second may disagree proposing that the object is much harder than was originally suggested. The soft-ontology approach, as described in 5.1, caters for both the above opinions. Each individual can put forward his or her suggestion on the object's hardness. Each opinion is processed as valid and taken into account when the domain is represented. The object, therefore, is classified according to the collective interpretation of its 'hardness'. The approach of soft-ontologies provides a manageable way to include the community in the process of organising their domain. Artefact properties are described by natural language. Moreover, by applying a scale to the term, soft is 0 and very hard is 1, individuals can choose a fraction to represent their interpretation of the object, such as the hardness of metal in swords or the lightness of metal in armour. The approach then utilises the soft-computing paradigm to represent the properties of a specific object.

In contrast, thesauri terms are rigid: a user may be unable to describe the softness they feel as indicative of a specific object. Objects are not necessarily recognisable as either hard or soft, to use the earlier example. Indeed the field of soft computing can illustrate uncertainty, which is often required when classifying content. It helps to

[54] Classification, as compared to categorisation, is the process whereby humans organise objects into specific fixed classes. According to Jacob, '*classification is the orderly and systematic assignment of each entity to one and only one class within a system of mutually exclusive and non-overlapping classes*' (Jacob 2004). Categorisation, however, is the cognitive process in which humans organise or divide the world into groups of similar entities.

involved a lengthy matching process[51] (Kilfeather, McAuley et al. 2003). The thesauri allowed users to classify digital content, in the form of narrative and other media, which was then submitted to the forum.

5.3. Aspects of the Soft-ontology model

This discussion compares and contrasts more traditional approaches to classification as instanced by the use of thesauri in the Explorer forum and more contemporary approaches such as the soft-ontology layer. The discussion, the author believes, is of particular significance with the rise in popularity of more collective approaches to classification, and therefore aims to investigate some differences between a top-down[52] versus a bottom-up[53] approach to indexing content. The chapter therefore considers research question, RQ1: *What are the difficulties of using traditional classification methodologies when approaching community-based platforms?*

[51] In order to use the English Heritage (EH) thesaurus of monument types, a mapping process had to be undertaken for each Irish monument type and archaeological object type to find where possible a matching terms within the EH thesaurus. The Irish classification system itself represents a resource of terminology, which describes the built heritage unique to Ireland. The mapping process therefore produced the following results. Of the 786 classes in the Irish vocabulary, 472 (60%) had a direct match to a term in the EH thesaurus. In this case the term was directly mapped. However 224 (28%) terms were closely related to terms in the EH ontology but were not linguistically similar enough to provide a direct mapping. In these cases the terms were mapped as preference terms, for example 'Ring fort'. A further 101 terms (13%) was used principally in the context of Irish folklore or archaeology and did not have a match with any English heritage terms. There is a further discussion outlining the work in Appendix B: Developing Explorer.

[52] In the context of the thesis, top-down methodologies (or approaches to organising content) refer to the creation of structured metadata models, such as thesauri and ontologies as discussed in 2.2, often developed by a specific group for use by a broader audience.

[53] Bottom-up methodologies relate to the social classification of digital content. Social tagging is the most popular name given to the practice by which the overarching community collaborate on organising and indexing content. Folksonomy, as opposed to taxonomy, is often the name given to the outcome of this practice.

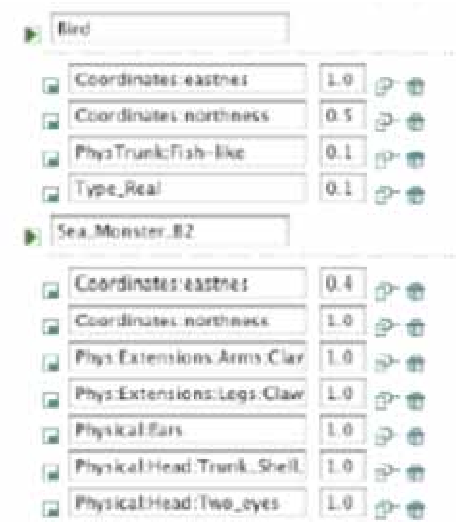
object vector. The object vector places the object within the overall domain. Similar object vectors are clustered together as illustrated in Figure 6.

When the process is complete and a vector is created for all domain objects, the SOM neural network (see section 2.3.1) can then used to cluster, represent and visualise the data.

5.2. The application of Standard Thesauri

The Explorer forum (section 1.4.1 and Appendix B: Developing Explorer) was developed by researchers at DIT with the goal of enabling broad communities of interest to record stories based upon Irish cultural heritage. To help classify contributions from the community, both in terms of narrative and related media, researchers working on the Explorer forum implemented several tools with thesauri support.

Thesauri, as discussed in section 2.2.1, are a recognised tool to help classify artefacts and disambiguate terms. As there was, at the time of development, no standard thesaurus or ontology representing the domain of Irish history, domain experts adopted and refined two English Heritage thesauri (discussed in section 2.2.1). The first was the Archaeological Objects thesaurus, which was developed to help identify portable evidence that resulted from past human activity. The second was the Monument Types thesaurus, which related types of monuments built and buried in England. The approach of researchers at DIT reflected the earlier work of Noy and McGuinness (section 2.2.3) in reusing existing structures and thereby standardising the domain, while, further, reducing the time associated with the engineering process. However, both thesauri needed to be refined and adapted to reflect the Irish domain. The work was undertaken by a member of the Discovery Programme staff and



Bird	
Coordinates:eastnes	1.0
Coordinates:northness	0.5
PhysTrunk:Fish-like	0.1
Type_Real	0.1
Sea_Monster_82	
Coordinates:eastnes	0.4
Coordinates:northness	1.0
Phys:Extensions:Arms:Clav	1.0
Phys:Extensions:Legs:Claw	1.0
Physical:Ears	1.0
Physical:Head:Trunk_Shell	1.0
Physical:Head:Two_eyes	1.0

Figure 7: Illustrates the soft ontology tool being used to describe monsters from the Carta Marina. The user in this illustration has entered the description of two artefacts. The first is a bird and the second is a sea monster. They have gone on to describe the various properties of the different objects, i.e. Bird - Type_Real and Sea_Monster – Physical:Head:Two_eyes. The weighting is to the right of each property.

The approach involves assigning each property an attribute vector through random vector encoding (RVE) techniques (RVE was originally developed by (Honkela T. et al. 1996) as a means of encoding large corpora of text for further processing by the SOM algorithm, as discussed in section 2.3.1). The technique produces a multi-dimensional vector of which the components are randomly chosen. The set of attribute vectors is used to describe every object within the domain. Each object is assigned a weight vector with the globally shared dimensionality that specifies the level to which each of the attributes characterises it. For every attribute the user estimates a weight variable that corresponds to the relevance of that particular attribute. The sum of the individual attribute weight vectors adds up to the overall weight vector for that object. The unique description of the object is computed as the object-specifically weighted sum of all of the attribute vectors and is known as the *using traditional classification methodologies when approaching community-based platforms?*

5.1. The application of Soft-Ontologies

The Nordic forum (section 1.4.2), which aimed to explore the narrative represented on the Carta Marina map, implemented several innovative techniques and practices to help organise and visualise the underlying domain content. One such approach involved the creation of soft-ontologies, an innovative way of organising a domain via soft-edged classes or a soft ontology layer (SOL).

A soft ontology layer (SOL), as described in (Collao, Diaz-Kommonen et al. 2003), provides a means of numerically categorising objects within a given domain. Researchers at UIAH developed a SOL tool that allowed users to manually enter a SOL of non-hierarchical properties and feature descriptions of a heritage artefact (see Figure 7). The process produces a low level, non-hierarchical ontology that compares objects by their properties. Unlike formal ontologies (see section 2.2.2), composed of non-overlapping concepts and their properties, soft-ontologies use natural-language definitions to describe object properties, as illustrated in Figure 7.

5. Community Interpretation

This chapter considers the introduction of soft-ontologies as a novel approach to organising and visualising digital content. Soft-ontologies were developed by researchers at UIAH and implemented as part of the Carta Marina forum (see section 1.4.2). The approach borrows from the work on SOM, as discussed in section 2.3.1, and proposes the creation of a soft-ontology layer to cluster similar artefacts in a specific domain[49]. To help contrast the application of soft-ontologies, the chapter will also consider a more traditional method of classification, that of thesauri[50], as implemented in the Explorer forum (see section 1.4.1).

Firstly, the chapter will introduce the notion of a soft-ontology and describe the process by which soft-ontologies are developed. Secondly, the chapter will introduce the approach of researchers at DIT when implementing two standard thesauri in the Explorer forum. The chapter will go on to discuss the application of soft-ontologies under the headings: collective interpretation, method of contribution, community models and limitations. However, each heading will further consider both the application of soft-ontologies and that of standard thesauri. The chapter concludes with a summary outlining some of the issues encountered by traditional techniques of classification, and the benefits of the soft-computing model. The research question that this chapter will attempt to answer is as follows, RQ1: *What are the difficulties of*

[49] The domain in this case was the Nordic Heritage represented on the Carta Marina and in 'A description of the Northern People'.

[50] Two English Heritage thesauri – the artefacts and monuments thesauri – were adopted and refined by researchers working on the Explorer forum. The thesauri were developed to help with major computerisation efforts ongoing in English Heritage. Both thesauri can be found at http://thesaurus.english-heritage.org.uk/frequentuser.htm

RQ3: *What impact can the creation and implementation of a structured ontology have when representing narrative concepts?*

4.4. Summary

This chapter introduced three research questions that emerged from the work presented in chapters' 2 and 3. Each question will help to structure the discourse in the remainder of this thesis. The questions surround, in one way or another, the complex issue of knowledge representation, ranging from the issue of hard-edged classification schemas to the use of narrative in an online community-based platform. The remaining questions specifically focus on community semantics, or the impact of the community on the process of domain representation and, later, the use of ontologies. The final research question, however, attempts to examine, in pragmatic terms, the advantages to both the use of simple and structured ontologies.

The next chapter will aim to investigate and answer research question RQ1. In doing so, the chapter will examine two contrasting approaches to classification. The first is the application of soft ontologies, an innovative approach developed by researchers at UIAH. The second is the application of a standard thesaurus, as implemented by researchers at DIT. The aim of the chapter, however, is to identify the problems with traditional approaches to classification when incorporating a community into the process.

asks: *What factors can influence the process of knowledge engineering when involving a community of non-technical users?*

4.3. Simple versus Structured Ontologies

Narrative, as discussed in 2.4, is central to how communities learn, develop and exchange knowledge. It was identified as a prominent tool for use in constructivist learning theory. There have been attempts by Schank amongst others to explore the possibilities of narrative through the application of new technology (see section 2.4.2). Indeed the field of narrative intelligence sprung up around this concept. Similarly, Landow suggested that new technology would produce a new approach to narrative, one free from the confinements of print text. This introduces the broader question of how to represent narrative in a virtual environment? Schank, in 2.4.1, suggested that human memory can be broken into two separate entities, semantic memory (a Memory for concepts) and episodic memory (a memory for stories). Semantic memory is very much like a conceptual model, which represents the most pertinent concepts of a specific episode. The episode or story, however, is held in episodic memory. In this way episodic memory provides the mechanism by which interesting stories are remembered, and the key concepts of the stories are retrieved from semantic memory. Taking Schank's approach to memory and recall, there are still, as identified in section 2.2, several ways to structure and represent narrative concepts. Ontologies, for example, provide a means to structure information semantically; nevertheless, there are several approaches to the use of ontologies. It is from this perspective that chapter 7 seeks to investigate the advantages to representing narrative concepts through the application of structured ontologies. Therefore the chapter inquires:

perspective, RQ1 asks: *What are the difficulties of using traditional classification methodologies when approaching community-based platforms?*

4.2. Community Definition

There are, as introduced in 3.2, several models of online community, and each tend to exhibit qualities of one or another at any one point in time. Similarly, there are several approaches to knowledge representation as identified in 2.2. The different community models suggest the information designer should, when undertaking the process of knowledge engineering, be cognizant of the participating community. Therefore the designer is not only concerned with the type of knowledge formalism but also the type of community, whether it is a CoP or a CoI for example. The type will firstly help define the formalism and secondly adopt and later refine the completed version. It is reasonable to expect different community models to benefit from different approaches to both knowledge engineering and method of contribution. For example, a tightly knit task-based group such as a CoP (section 3.2), who sometimes interact offline, may be able to develop the community's domain ontology and later create and manage online content without the need for predetermined moderators (section 3.3). Alternatively, a broader community, or CoI for example, may require some vetting of content to ensure that contributions are of an appropriate nature. This introduces the broader question of community platform, and indeed social policies, as discussed in 3.3. Within this context, both Clark and Preece used the analogy of a community designer with that of the mayor of a town, who organises governing policies and develops a suitable infrastructure to help members of the community strive as a satisfied unit. However, concentrating on the community definition, and the impact of a community on formally structured knowledge, RQ2

4. Discussion: Proposed Research Questions

This chapter presents a discussion on several of the topics and approaches that were set forth in the previous two chapters. The chapter attempts to bridge the disciplines of knowledge representation and online communities, and in this propose several questions to help structure the discourse of the remainder of this thesis. The chapter is therefore divided into three parts, each presenting a discussion with the emergence of a question that is approached in a later chapter.

4.1. Community Interpretation

Ontologies, according to both Fensel and Gruber, are developed upon a *shared understanding* (section 2.2). This indicates involvement on behalf of the community. But how is a community involved in the engineering process? This thesis is not concerned with large-scale ontology engineering. If an institute or engineering body, for example, has a large number of engineers to collaborate on a specific ontology, one which is based in their sphere of knowledge that is, then it could be, albeit rather simply, assumed that increasing the effort mitigates the problems arising from the engineering process. However, online communities, as discussed in chapter 3, are not generally made up of groups of engineers. CoIs, for example, tend to exhibit a broader demographic, and are not, in the context of this thesis, concerned with engineering. Nevertheless, a key role of the information designer is to empower the community to enhance or develop their own world-view. Standard vocabularies, such as those discussed in section 2.2.1, introduce a priori taxonomy, which can lie outside the community's vocabulary. Standard vocabularies are often fixed, immutable and, for the most part, do not change or adapt as the community develops. However, the nature of social platforms is to reach, involve and include the community. From this predicated on the completion of a task, such as CoPs, others come together to develop a collective knowledge base, as exhibited by Wikipedia, while others again develop friendships and relationships online, as illustrated by the vast number of popular social networking sites.

The chapter concluded with a discussion on some of the more traditional aspects of online community design. Information, or OC, designers, in this context, were compared with the mayor of a town who develops an appropriate infrastructure and sound polices to help the community thrive as a whole. In this way the community is grown, not built, and the job of the designer is to identify polices and technologies appropriate to each specific community. It emerged from the chapter that although there are several models of online community in existence, several overarching guidelines, such as purpose, policies and technologies, could help to sustain a successful online community.

The next chapter will present a discussion based on many of the concepts and approaches reviewed in the last two chapters. The discussion will aim to identify several questions, which emerged from the review of literature, and that help to structure the remainder of this thesis. The questions will focus on approaches to representing knowledge but will further consider the impact of a non-technical community or, as identified in 3.2, a CoI or CoP, on the collective approach to knowledge representation.

Kollock's sources are not concerned with online communities many principles that evident in traditional communities carry over to the virtual world.

Amy Jo Kim, author of *Community Building on the Web* (Kim 2000), describes three questions unearthed from her own personal experiences that any online community designers should ask themselves:

- What type of community am I building?
- Why am I building it?
- Who am I building it for?

These questions help define a community's purpose. She identifies 'purpose' as an essential component of the community model, maintaining that it provides the community with a *raison d'être* and helps to keep the community members focused. She goes on to suggest the importance of leadership, as instanced by Clark previously and the significance of roles, which, she maintains, couples power, i.e. moderator-ship with responsibility, must be provided to include newcomers without alienating regular participants. Taking a brief look at traditional approaches to online community design, it could be argued that purpose, identity, policies and platform emerge as pertinent to a successful online community.

3.4. Summary

This chapter introduced the concept of a virtual community. Originally online communities were thought of as a marginal communication paradigm but over the past decade or so they have evolved to encompass entire knowledge and social platforms. As discussed in 3.2, this evolution has engendered the birth of several different strains of what originally represented an online community. Some are which users can communicate and build substantial relationships throughout an online community (Preece 2000).

The ability of members to identify each other helps build *social capital* amongst the group, while a history of a user's previous postings provides an overview of that individual's online persona. Slashdot[48], the technology community website, encourages users for identified contributions through the use of the term '*Anonymous Coward*'. Amy Bruckman, in her article '*Finding One's Own in Cyberspace*', contends that members who are unlikely to share their true identity are equally unlikely to engage in serious discussion (Bruckman 1996). Similarly, as mentioned in 3.2, problems associated with anonymous contributions have resulted in Larry Sanger's project, Citizendium, which favours a smaller community with the use of identity to avoid amateurish contribution. Thus, it would seem that when dealing with professional communities, or communities striving towards a common purpose, identity is essential for successful and purposeful contribution.

Kollock, in his paper *Design Principles for Online Communities*, mirrors the views of both Bruckman and Sanger. He identifies three conditions that are necessary for *even* the possibility of cooperation: firstly, ongoing interaction and communication, secondly, identity and the ability of individuals to identify each other and, thirdly, a history of member's previous behaviour. He stresses the importance for each group to customise the norms and rules that governed their behaviour and members sanction and monitor member's own behaviour (Kollock 1996). Although

[48] The technology web site, http://www.slashdot.org, operates on similar principles to Digg mentioned in the introduction to this thesis but concentrates on subjects surrounding technology. Most of the contributors champion open source software.

more likely to complete a given task than a scattered or unfocused CoI. However, this section does not intend to differentiate between community models. Here factors that help promote successful online communities are discussed.

Clark maintains that the most successful online communities are not built, but grown. He states that one can create the environment and plant the seed but it is the members themselves who grow the community (Clark 1998). He further states that: '*the virtual environment is a mediated environment and mediation needs mediators*', just as the WELL has its hosts and listservs[45] have list-moms[46], online communities need to have leaders. Preece follows Clark's analogy by comparing the community developer to the mayor of a town '*who works with town planners to set up suitable housing, roads, public buildings, and parks, and with governors and lawyers to determine local policies*' (Preece 2000). Although there is no one solution to community development, there are, however, a number of design elements that can have an important impact on encouraging and supporting a successful online community. From this perspective, Preece highlights two principle aspects[47] that impact the nature and success of an online community. They are, firstly, the method of communication and, secondly, what Preece terms as '*sociability*', or the ease at

[45] Listserv is an electronic mailing list software application developed during the 1980s. The term 'Listserv' has become synonymous for all mailing list applications although the original was developed Bitnet computer network

[46] List-moms are the term used to describe people who mediate listserv forums.

[47] Similarly, albeit more recently, Clay Shirky identifies three rules that underpin the success of social software. He suggests that for social software to prove successful it must rely on '*a successful fusion of a plausible promise, an effective tool and an acceptable bargain with the users*' (Shirky 2008). Interestingly, his comments reflect those of Preece's in developing an environment that can accommodate social interaction and develop a community. However, with the rise of the social web and the increasing number of social tools, Shirky introduces both promise and bargain to attract and sustain groups of users in an increasingly competitive market place.

not strictly sticking to a single definition. For this reason, the different models have been discussed in this chapter.

As discussed in this section, the popularity of online communities has created a diverse number of community models due to the differing requirements of the user base. However, the success of online communities is predicated on a number of differentiating factors such as communication platform, and the social policies governing the group. The following section groups these factors under the heading of traditional support for online communities where they are discussed in greater detail.

3.3. Traditional support for online communities

Within the increasingly large body of literature concerning online communities, it is clear that there are a variety of factors, which contribute to the successful launch, and evolution of an online community. Although these factors add to a community's success, there is no definitive way to develop an OC. It was difficult, for example, to foresee the meteoric rise of BeBo, Facebook, or indeed Youtube. Those websites have, however, indicated that users are more than willing to contribute knowledge and time for the greater good of the community, as further evinced by Wikipedia. Furthermore, it has also become clear that users are also willing to share and publish content, make lengthy or exhausting online profiles and spend hours communicating online. The current generation of users are growing up with concepts such as social networking, a phenomenon that, at time of writing, helped make Facebook the most popular website on the planet, overtaking both Yahoo and Google, the most prevalent names in cyberspace. As has been discussed in the last section, there are several community models (discussed further in the next couple of chapters) some are more inclined to certain approaches than others, for example, task based communities are belonging to other communities during any phase of its existence, or possibly pass through these categories independently (Carotenuto, Etienne et al. 1999). For this reason, CoIs are here defined as being '*composed of people who typically share common backgrounds or interests*' (Carotenuto, Etienne et al. 1999). Examples of communities of interest include a group of citizens concerned with the preservation of forestry, the WELL or members of a community interested in local heritage.

From the perspective of community metamorphosing between models, Baxter defines three other varieties of online community (Baxter 2002). Communities of circumstance are groups of individuals brought together through some common situation such as the Dublin City Collective[43], an online community for people who live and work in and around Dublin city, Ireland. Communities of purpose are groups of individuals, who have a tighter focus on a common interest. Members of a community of purpose often come from a wider range of backgrounds than a community of practice. To this end, they are less likely to have a deeply shared view of the domain but share a deep conviction for the success of the enterprise (Carotenuto, Etienne et al. 1999). Communities of purpose, such as the knowledge management community at the Knowledge Board[44], have members from numerous disciplines concentrating on the KM field. Finally Baxter identifies Corporate Communities as business communities, which develop to strengthen both internal and external relationships, harvest knowledge and develop a 'corporate culture'. Although these community models are not directly related to this thesis, as stated by Carotenuto, community models often exhibit traits in a more heterogeneous manner,

[43] The community of circumstance, Dublin City Collective, can be found at dublin.citycollective.com.

[44] The knowledge board is available at www.knowledgeboard.com.

and today Wikipedia contains millions of articles in over one hundred different languages.

Yet the model has come under recent criticism from long-term Wikipedia expert Larry Sanger. He suggests that the widespread anonymity of Wikipedia has led to several problems: community members are unwilling to enforce their own rules, leaders are becoming increasingly insular and new members are finding it progressively difficult to become involved in the community (Sanger 2006). As a result, he is in the process of creating an alternative to Wikipedia, a '*progressive fork*' called Citizendium[42], which advocates '*public participation with gentle guidance*'. Sanger proposes that with Citizendium, a large community of intellectuals could come together, in a similar fashion to that of the open source community, and collaboratively strive towards a community based online compendium.

Supporting effective and progressive collaboration has spurred the creation of, what Diaz-Kommonen describes as '*ephemeral communities'*, communities that come into existence for the purpose of engaging in an activity for the duration of a given task (Diaz-Kommonen 2002). Ephemeral communities do not, necessarily, engage in a virtual environment. They are transitory, and therefore can become inharmonious over periods without active engagement or discontinuity. Nevertheless, they are flexible, expandable and invaluable as a collaborative intermediary in the activity of multidisciplinary design.

Aside from ephemeral communities, more traditional online communities models are not mutually exclusive, and it is suggested that any community can exhibit traits

[42] Langer Sanger's alternative to the Wiki model may be found at http://www.citizendium.org.

specific task orientated purpose, CoIs are often positioned to support innovative practices and evolve as a community model.

The WIKI phenomenon gave birth to another type of community which is not necessarily represented by CoPs or CoIs. A WIKI community exists on the periphery of traditional community models, as the identity of a contributor is often never revealed. Identity, as will be discussed later, is proposed as fundamental to the success of traditional online communities. By contrast advocates of WIKI communities propose the model as inherently democratic, as each member is encouraged to contribute unimpeded by familiar biases, which traditionally occurred in the face-to-face collaboration. The problem of anonymous contribution has always been an issue within the online community sphere. Even the earliest systems, such as Delphi[40], tackled the difficulties of physical collaboration by introducing different degrees of anonymity. Both Turoff and Hiltz, who worked directly on the system, understood that often people are unwilling to commit themselves to initial expressions as they may turn out inappropriate, or similarly people of eminent status are cautious at introducing questionable ideas as they may undermine their strong position within a group (Turoff and Hiltz 1996). Later CMC[41] presented a way to circumvent issues, which hinder progress in face-to-face group processes. The success of the anonymity model is evident from the large number of WIKI communities scattered across the web. By far, the most prevalent of which is Wikipedia, an online encyclopaedia founded on the principles of open contribution. Consequently content mushroomed

[40] The Delphi system was proposed in 1975 by Linstone and Turoff as a way to structure and organise group communication. The principle lay with the ability of software to allow the individual and the group as a whole to deal with complex problems (Turoff and Hiltz 1996).

[41] CMC is an acronym for Computer Mediated Communication, which can be used to identify any a mean of communication enabled by computer networks.

specific query. Examples of such communities include Sun Corporation's Java Programming Forums[38] and the Experts Exchange forum[39]. This form of community is very popular with people working in technical disciplines as it provides access to a large and diverse knowledge base of similar professionals. It is also a reflection of the origins of the medium, where technical professionals congregated online to make use of a collective knowledge base.

In contrast to both networks of practice and Lave and Wenger's CoPs, CoIs have often a looser focus but are united by a universal interest. Fischer maintains that CoIs draw together individuals from different disciplines, and therefore CoIs may be thought of as heterogeneous communities whose members or stakeholders come from diverse CoP (Fischer 2001). This is because often members involved in a community of interest are contributing solely out of personal curiosity, rather than professional inclination. Nevertheless, personal curiosity often leads to a more developed sense of belonging. In such cases, members may emerge as stakeholders, of community champions who have key roles or greater responsibilities in the community. Nevertheless, CoIs often include members from a range of separate disciplines but their professions may not contribute to the community's forum. Within this context, CoIs are looser models of community participation in that they are not constrained by the institutional practices that often bind CoPs together. Without the obligation of a

[38] There are several other technology based online communities where users contribute and glean knowledge. Expert's Exchange, found at www.experts-exchange.com, provides a broad resource in that no specific language is targeted; however, access is provided for a small fee.

[39] The Java.sun.com website provides several resources for those wishing to learn the Java language. There are several detailed community forums targeting specific aspects of the Java programming language.

Hung & Nichani maintain that '*a distinct characteristic of enculturation within communities of practice is that of learning to be*' (Hung and Nichani 2002). However, discourse, which takes place within online communities, is about learning something. Contrary to Lave & Wenger's original conception of a CoP, participants of Internet communities are involved in discourse about knowledge, rather than explicitly learning to be through the engagement of the community processes. It can be argued that CoPs, therefore, cannot exist solely online, that a virtual platform can enable long-distance collaboration but that understanding more implicit forms of knowledge can only be achieved through physical participation within the community processes.

To distinguish physical and virtual learning communities, Brown & Duguid, in '*A social life of information*', define a second type of work-related community as a network of practice (Brown and Duguid 2002). A network of practice is a system that links people to others whom they may never get to know but who work on similar practices and share comparable knowledge. These types of communities are defined by Hung & Nichani as quasi-communities, in that they support a question-driven forum facilitating learning as a by-product of '*a need for wanting something*' (Hung and Nichani 2002). Communication is carried out indirectly through, for example, newsletters, web sites, discussion forums and coordination as a result is quite explicit. What unites these networks is the term *practice*. Such networks are also called '*social worlds*' or '*occupational groups*'. Participation in networks of practice is usually based on specific needs and demands where members contribute because there is a rather quick way of receiving some benefit; for example, other members answering a engaging with a wider audience. This of course was aided through the provision of increasingly inexpensive communication tools.

Certain groups, however, do not fall into the category of CoI. Some groups might be considered 'learning communities' and their members do not join due to a common interest, but rather to learn, participate or for the completion of a specific task. Such communities are sometimes known as Communities of Practice (CoP). CoPs, according to Wenger et al. *'are groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis'* (Wenger, McDermott et al. 2002). They argue that learning is not something an individual will do, but rather is a social activity and comes from experience within collective life. In this way, learning is not thought of as a passive exercise; rather it involves active participation, developing the character, behaviour and outlook of the practitioners. Hence, people with similar professional interests or requirements organise themselves into homogeneous communities, learning knowledge from other members within the group.

Lave and Wenger reinforced their concept of a CoP with the introduction of Legitimate Peripheral Participation (LPP) (Lave and Wenger 1991). LPP is often thought of as a simple apprenticeship model where the master confers knowledge on the student through situated learning, i.e. the student is embedded in a social and physical environment, yet the concept of LPP is not solely restricted to this mode of learning. Lave and Wenger contend that within a community of practice, newcomers learn from established members and in time become established members themselves: learning is thus seen as an integral part of the practice (Hildreth, Kimble et al. 2000). However, Lave and Wenger were not focusing on online participation. In this context, it is often argued that CoPs cannot survive solely through a virtual platform.

caring, gardening and software development are all examples of areas and practices that recognised the benefits of a community model contributing to a more successful platform. As the ubiquity of OCs became increasingly evident, several distinct types of community began to emerge, categorised by their purpose, shared characteristics and platform of use.

3.2. Online Community Models

Initially, online communities were classified as Communities of Interest (CoI) (Fischer 2001), where likeminded participants engaged in a discourse surrounding a common purpose or concern. These groups were small and few. The publication of Howard Rheingold's book '*The Virtual Community*' (Rheingold 1998), however, helped to catapult the concept of an online community into the mainstream. His story involved the WELL[35], a place for likeminded people to form strong friendships and even find lifelong partners. Since the publication of 'The Virtual Community', thousands of CoIs[36] have developed around every conceivable interest[37]. As the popularity and success of the online community paradigm increased, the functionality, purpose and demography of users helped to shape the community model as a whole. Therefore, what previously might have been considered as a means of reaching a specific or likeminded community was suddenly recognised as a successful method of

[35] Rheingold's book detailed life within the well's virtual community. The well, available at www.well.com/, became a space for users to build deep and lasting relationships as outlined in his book, The Virtual Community (Rheingold 1998).

[36] Examples of general purpose CoIs are www.communities.com, www.boards.ie, groups.yahoo.com and groups.google.com where a member may begin their own community pertaining to a particular interest.

[37] It is worth mentioning that although the WELL can now be categorised as a COI, it was in only subsequent years, following the success of the WELL and the publication of Rheingold's book, did the notion of a COI, and that of a COP, begin to emerge. Having said this, the WELL fits the definition of a COI, as discussed by Fischer in 3.2, and consequently the author uses this definition, albeit retrospectively, to identify with the activities of this popular online community.

internet known as 'ARPANET[34]'. The nineteen eighties saw the development of early online communities by groups of users with similar goals and experiences using similar communicating software - bulletin boards (Preece, Maloney-Krichmar et al. 2003). However, towards the latter half of the nineties, an unprecedented increase in online community participation was observed. This increase is described in a 2001 Pew Internet & American Life Project report cited by Preece, which states that '84% of all Internet users indicated that they contacted an online community and 79% identified at least one group with which they maintained regular online contact' (Preece and Maloney-Krichmar 2003).

Although early online communities were regarded as a social phenomenon, the significance of the Internet within the commercial sector and the economic value of the new communication paradigm quickly became evident (Stanoevska-Slabeva and Schmid 2001). With the business sector taking notice, a shift in online-community thinking took place: no longer were OCs solely a place to meet and socialise with like-minded people. Service providers, such as eBay, saw the benefits of community participation, and provided collaborative auctioning tools to support the exchange of goods. With eBay, however, the community was thought of as a means to an end and not an end in itself. Retailers recognised the advantages of user-created content and developed services such as Amazon's review community, allowing users to contribute their personal critiques about products and services on the Amazon website. It was not just the commercial sector that began harnessing the innate power of the internet through online communities: education, knowledge-management, engineering, child-

[34] Arpanet was the first packet-switching network proposed by DARPA, Defence Advanced Research Projects Agency, in the late 1960s and developed into a successful prototype and working Internet in the late to middle seventies (Leiner, Cerf et al. 2003).

3. Online Communities

This chapter introduces the concept of an online community. The chapter is broken into three distinct parts. The first discusses the evolution of online communities, from the early days of ARPANET to the more contemporary or ubiquitous social platforms. The second examines the different models of online community that have evolved over the past number of years. This is important, or so the author will later argue, because the type of community, whether a CoP or CoI for example, can heavily impact the choice of knowledge representation or indeed the entire approach to representing a community's domain of knowledge. Within this context, it is important to understand the dynamics and requirements of specific community models. The third part investigates some of the more successful approaches that have cultivated online communities in the past. Although 'web 2.0' has spawned a variety of platforms founded upon community contribution, these platforms are still, predominantly, developed upon asynchronous technology. Therefore many of the technology platforms, social norms and policies that once helped to coordinate groups and encourage collaboration are still relevant today.

3.1. Evolution of online communities

Historically, there has been great progress in point-to-point communications (telegraph, telephone, mail, fax…) and in broadcast technologies (mail, radio, television, books….), but parallel progress has not evolved in the sphere of group communications. Since the development of the Internet, however, communities of users, although geographically remote, are able to collectively communicate through this new medium. Initially, online communities grew through the use of computer networks on college campuses and amongst engineers working on the precursor of the

The chapter then presented an approach to automatically creating semantic representations based on some prior input. SOM (section 2.3.1), developed by the Finnish professor, Teuvo Kohonen, clusters entities according to their similarity. Unlike the approaches discussed under the heading of Ontologies, however, SOM reflects the soft-computing paradigm and, when applied, produces representations composed of soft-edged classes.

The chapter concluded with a discussion on narrative. Narrative is, as proposed by the author and academic David Lodge, one of the principle ways in which humans relate experience and transfer knowledge. It is integral to how we learn and draws close parallels with both the social and psychological strands of constructivist learning theory. For many, however, the introduction of new technology suggested innovative ways of approaching the narrative form. Section 2.4.2 outlined the ideas of Landow on hypertext fiction, where he proposed an end to the confines of traditional print text by, enabling a paradigm shift, away from the centre, margin, and hierarchy. The reader, Landow suggested, could experience a greater sense of immersion, agency and transformation. Yet, hypertext fiction failed to capture the publics' imagination, (as Landow and other post-structurlists once thought) as the public, it seems, are more engaged with traditional forms of storytelling.

In the next chapter, the discussion will explore the notion of an online community; a concept, which, in light of more contemporary social platforms, has certainly captured the public's imagination. However, the chapter will not concentrate on any one specific platform or community, but rather discuss a range of models considered under the rubric of online community. The discussion will further examine how these communities are brought into existence, function successfully and are supported with effective social policies and technology platforms.

So although new technology proposed new and exciting ways of interacting with narrative, we have not seen a proliferation of hypertext fiction. Users, or readers in this case, have instead opted for traditional approaches to narrative, preferring the linear over the non linear. New technology, however, is arguably most predominant in the social or everyday activity of people, from mobile phones to instant messaging, email and the online community. If narrative is how we relate, share experience and transfer knowledge, technology is often the facilitator. The next chapter therefore returns to the basis for this thesis, the online community, but concentrates on the models of community that have emerged over the past decade or so and concludes with a discussion on how online communities have been traditionally supported.

2.5. Summary

This chapter introduced the discipline of knowledge representation. Firstly, ontologies were discussed as an increasingly popular way to represent and structure real-world knowledge. This popularity, it could be argued, stems from the growing interest in Tim Berner's-Lee's semantic web vision. Nevertheless, and as was identified in 2.2, there are several approaches to both the development and use of knowledge models. Simple ontologies, for example, comprise of natural language terms, while more structured ontologies consist of concepts, and their inter-relationships, expressed in a formal language. Essentially, the difference between both approaches is based on the ability of a machine to reason across formal and explicitly stated knowledge. Simple ontologies do not express knowledge in this way. It is within this context that Fensel, when discussing the attributes of a formal ontology, states that '*the types of concepts used, whether explicit or tacit in a real-world sense, should be explicitly defined*'.

Procedural authorship means writing the rules by which the text appear as well as writing the texts themselves' (Murray 1997). To this end, she contends that with hypertext narrative a reader does not experience authorship but a deeper sense of agency.

Landow maintains that hypertext instantiates another quality of Barthes's ideal text by blurring the boundaries between the reader and the writer (Landow 1991). This can be demonstrated by the lack of closure found in hypertext narrative. Closure occurs not when the plot is finished or run-out but when the reader decides. This refusal of closure supports the third aspect of the medium, *transformation*. The reader is no longer constricted by time as with the linear form of narrative. Instead they are afforded the opportunity to change character and explore the narrative space from different perspectives. The reader may shift from the central protagonist to another character within the same story. The ability to alter characters supports the construction of a composite view of the narrative world. The single story may be constructed into a coherent structure of interrelated narratives. However, the effectiveness in which this form of narrative has been implemented is still questionable, as people still opt for the more traditional forms of storytelling. A fictional narrative may offer the reader alternate view points of the central narrative, switching between the characters much like multiform stories in more traditional methods of storytelling. Furthermore, with non-fiction, the reader may experience a turn of events from another character's viewpoint. However, control offered to the reader is usually severely limited as it may alter the coherence of the story. Moreover, readers are often left dissatisfied as they are left with the feeling that another path could have provided a more interesting and satisfying conclusion.

regularly used when a reader finds a narrative particularly enthralling. New technology offers the potential to increase this effect of immersion by engaging the users through augmented interactivity with the narrative. Immersion can lead to engagement. Douglas and Hargadon suggest that the transition from immersion to engagement may frustrate many readers and can lead to a feeling of disorientation. Disorientation may be compared to '*being lost*' and is partly due to the readers awareness that hypertext exists in a virtual space. While technology facilitates a greater sense of immersion, the author must be careful to create a coherent experience for the reader (Douglas and Hargadon 2001).

The second aspect of hypertext narrative, which needs to be discussed, is *agency*. Agency may be defined as the ability of the reader to take action and see the immediate results of that action. Computers, as reactive machines, naturally exhibit agency: opening a file or typing words to the screen are both examples of agency. Hypertext fiction, such as *Victory Garden*, allows the reader to determine the outcome of the story by choosing a path through the narrative. Post-modern hypertext tradition takes the undetermined text out of the hands of the author and places it into the hands of the reader (Landow 1991). This shift creates a heightened sense of agency for the reader, as now it is he who decides the outcome of the story. Michael Joyce enforces the sense of agency in the interactive fiction, *Afternoon*, by telling the readers to decide for themselves when the story is finished. He titles it a '*work in progress*' and states: '*Closure is, in any fiction, a suspect quality, although here it is made manifest. When the story no longer progresses, or when it cycles or when you tire of the paths, the experience of reading it ends*' (Murray 1997). This shift has led to the argument that the reader or 'inter-actor' becomes the author of such digital narratives. However, Murray argues that '*Authorship in electronic media is procedural.*

It was felt, however, by many theorists that technology would pave the way for new and remarkable narrative possibilities, as the realisation of hypertext broke the confines imposed by print text. Although the idea of '*hypertext*' is older than the modern electronic computer (Bush 1945)[32], many critics dismissed its narrative possibilities while hypertext enthusiasts emphasised the pleasures of increased interactivity. George P. Landow, an evangelist of the medium, argued in his book *Hypertext: the Convergence of Contemporary Critical Theory and Technology* that technology enables a paradigm shift, away from centre, margin, hierarchy, and linearity toward new post-structuralist narrative forms (Landow 1991). The possibilities available through hypertext narrative were explored in such pieces as Michael Joyce's *Afternoon* and Stuart Moulthrop's *Victory garden*[33]. Landow, in illustrating both Barthes' and Derrida's philosophies on hypertext, proffers the perpetually unfinished textuality of a hypertext system described in terms of links and nodes of lexica (Landow 1991). He describes how the new medium allows the author to break the chains imposed by traditional linear narrative. However, the wonderful possibilities being proposed by the proponents of hypertext narrative systems lay principally with the aesthetics of the medium; immersion, agency and transformation.

Narrative immersion is a metaphorical term derived from the physical experience of being submerged in water. In relation to stories, immersion gives the reader the impression of being surrounded by the narrative (Murray 1997). The immersive power of narrative is not a new phenomenon: '*being lost in a book*' is a phrase

[32] Bush's essay As We May Think, discussed the possibilities of making a collected human knowledge more accessible. In his highly cited and forward thinking essay, Bush first mentioned the possibility of a hypertext system called the memex (Bush 1945).

[33] Victory garden can be found at http://www.eastgate.com/VG/VGStart.html

that, although the following section, new technology, deals strictly with formal narrative, it does present many of the supposed advantages of writing for a new platform, i.e. hypertext. Furthermore, while many of the approaches discussed in chapter 7, concentrate on formal narrative, the author suggests that each approach can be applied to any form of written narrative - be it a protracted essay or diminutive missive. Each form can be a valid and worthwhile contribution to the community's knowledge base.

The following examines what has become known as hypertext fiction, a form of narrative that exploits the non-linear aspects of hypertext. It is included because narrative created for a hypertext platform is different from narrative written for print or traditional methods of publishing. Furthermore, hypertext fiction was an attempt to exploit the hypertext paradigm and provide the reader with a new way of interacting with narrative. Some of the approaches examined in chapter 7 attempts to use ontologies to drive or present narrative in a similar way to the early efforts at hypertext fiction.

2.4.2. Narrative and New technology

Narrative has traditionally been an area of much interest for AI researchers. Schank's group at Yale, for example, explored how the human mind processed knowledge structures while understanding the meaning of natural language (Mateas and Sengers 1999). Their approach resulted in the SAM (story-understanding) system and a theory on the sort of knowledge structures needed to understand textual narrative. Much of the work around this time focused on developing systems that could effectively generate or consume narrative. Nevertheless, an important aspect of this work concentrated on the representation of stories.

individually, but constructed through group collaboration and community participation.

Online communities have provided a similar function for people who, like Orr's study, work in a technical discipline but, unlike Orr's study, are not explicitly involved in a group or community of other professionals. The open source community, for example, provides free software and support through numerous message and bulletin board systems. Participation in this sort of community, it could be argued, does not contribute to learning as attributed by the social constructivists. The activity of a technical online community, however, involves posting questions or answers, or either finding or receiving appropriate or helpful answers. This does not reflect Vygotsky's theory of social learning and may propose a new model of learning predicated on virtual environments. Chapter 3, online communities, will discuss social learning and the online community paradigm in more detail.

Nevertheless, if narrative is central to the dissemination of knowledge through the everyday activities of a community (as suggested by Orr), how does this manifest in an online environment? It is important, at this point, to make a distinction between what the author identifies as formal narrative - created for a broad audience and, possibly, in the form of a personal essay, and informal narrative - created with a smaller circulation in mind, within a small group of friends perhaps, and in the form of an email or message board posting. The thesis will primarily focus on formal narrative - created by an author for dissemination within a broader community model. This model of activity is akin to the Wikipedia model of contribution. However, ignoring informal narrative is, in many ways, ignoring the vast mountains of knowledge that lie in the message board postings and Usenet emails that have built up over the past decade of online community activity. Therefore, the author proposes memory, while the key concepts such as the *chevaux de frise*[31] would be stored in semantic memory.

There are, however, two strands of constructivism: psychological/cognitive constructivism and socio-cultural/social constructivism. Both stem from the original concept of constructing meaning and they share many commonalties, such as the active participation of students. Psychological constructivism focuses on the way that students construct knowledge, while social constructivists, such as Vygotsky, place more emphasis on social factors when constructing knowledge and understanding. The use of narrative and reflective dialogue supports Vygotsky's theory of social learning and is demonstrated in Lave & Wenger's study on communities of practice and Legitimate Peripheral Participation (LPP) (Lave and Wenger 1991). This concept, of utilising the ubiquitous nature of narrative in supporting community learning, has been adopted by many corporations and organisations and has parallels in the field of knowledge management. One of the most widely cited studies is that of Julian Orr, an anthropologist who worked for the Xerox Corporation. In his book, *Talking about machines*, Orr describes how stories provided what he characterises as the 'perfect vehicle' for knowledge sharing and dissemination amongst Xerox technical representatives (reps) (Orr 1996). Orr's study showed that the natural way in which reps communicated and solved problems was through narrative. Reps would meet and discuss particular problems they had encountered and solved during their day's work. These casual meetings were how the reps shared their knowledge and experience. Learning in this context is seen as a social activity not undertaken

[31] In the above example the term Chevaux de frise represents sharp, potentially-lethal, upright wooden spikes designed to impede assault and found around forts such as Dún Aonghasa.

the constructivist hypothesis, reinforces Hein's view by highlighting the importance of learning from doing; he maintains that the accumulation of abstract facts is not learning, and will be forgotten in due course, while learning from doing and practice will remain with the learner throughout their lifetime (Schank 1995).

In accepting the constructivist's epistemological position, however, we have to, as Hein states, recognise that '*there is no knowledge "out there" independent of the knower but only knowledge that we construct for ourselves as we learn*' (Hein 1991). Constructivist theory questions an all encompassing descriptive knowledge, where the individual learner does not feature, and emphasises the individual, who as a human being creates his or her own interpretation of the world about them. Within this context, Schank argues that it is through narrative that humans achieve this mental representation of their surrounding world. He believes that human memory can be broken down into two separate entities, semantic memory (a Memory for concepts) and episodic memory (a memory for stories). Semantic memory may be thought of as an abstract conceptual model, for example, a student writing an essay about a field trip to the famous Irish hill fort Dún Aonghasa on the Aran Islands, may use semantic memory to represent the key concepts of the trip, the hill fort's architecture, the different occupants, the location and advantages for defence. Stories are placed in episodic memory while the knowledge gleaned from the story is represented in semantic memory. Episodic memory provides the mechanism by which interesting stories are remembered, and the key concepts of the stories are retrieved from semantic memory (Mulholland and Collins 2002). For example, a story about what is known about the occupants of Dún Aonghasa would be represented in episodic rather '*manifests in a discernible organisation*'. In this way, a story may comprise of several events ordered in a hierarchical fashion. While events may be considered as either acts or actions in which an existent is the agent of the event, in turn, an existent in this sense is either a character, or a setting (Chatman 1978). There is, of course, the matter of time, and Chatman makes the distinction between story-time, as time passing within the story itself, and discourse-time, as the time associated with telling the story.

It is clear from both Brooks and Chatman that narrative may be thought of as a structural framework consisting of story and discourse. A Story is the conceptual organisation of objects, events and individuals, while Discourse is the manner in which a story is told. The organisational and abstract structuring of narrative offers the poet, author, or storyteller a conceptual canvas on which to share life experiences. Its function, however, extends into the field of contemporary pedagogy, as narrative is now seen as having a pivotal role in the process of learning (Mulholland and Collins 2002).

2.4.1. Narrative and Learning

It is often argued that when viewed from a pedagogical perspective, narrative draws close parallels with constructivist learning theory. The constructivist hypothesis is predicated on the idea that everybody constructs their own personal meaning (either socially or individually) from the world about them. Hein, in illustrating the constructivist perspective, describes the act of learning as being primarily the learner's; he maintains that pedagogy does not solely involve learning and stating facts but rather the learner must be engaged to construct meaning and context from exhibits or lessons (Hein 1991). Roger Schank, another proponent of *peculiar to and universal among human beings*' (Lodge 1990). He maintains that narrative provides people with the solution to the problem of transferring knowing into telling; through narrative people can translate human experience into structured meaning. Its importance, however, lies not just in the telling, but with the re-telling. Narrative[30] provides a vehicle for the dissemination of knowledge throughout a community. It is central to learning, education and how people construct and share knowledge (Brown and Duguid 2002).

There are many interpretations of the terms *narrative* and *story*, particularly when discussed from a multidisciplinary perspective. To distinguish between the two, Brooks defines narrative as the '*manner of expression*' or the way in which a story is told, while a story is defined as '*a system of associations between elements, composed of events, people and things*'. He goes on to explain the relationship between story and narrative in this way: '*narrative represents the universe of story elements for a given story – the collection of possibilities – while narration represents a specific navigation through that universe*' (Brooks 1997).

Chatman, in his book *Story and Discourse: Narrative Structure in Fiction and Film,* seeks to tie down the elements of narrative theory. He suggests that a story is what is told, while a narrative is how it is told, but delves deeper into structuarlist theory by suggesting that a story consists of a chain of events (or actions, happenings) and, what might be termed, existents (or characters, settings, etc) (Chatman 1978). Referencing the work of Jean Piaget, Chatman suggests that narrative has a definite structure, and is not simply an agglomeration of events in an unordered manner, but

[30] See Appendix A: Narrative, collecting and creativity for a further discussion of narrative's relationship to context and curatorship.

SOM can also produce visualisations (both the functions of clustering and visualisation are illustrated in Figure 6). Visualisation provides the possibility of comprehending huge data sets, thus reducing the time needed to understand the information and reveal object relations that otherwise might have gone unnoticed. Therefore, via SOM's visualisation techniques it is easier to detect isolated patterns and structures within the dataset. This technique contrasts with conventional data retrieval methods that require the user to either have some prior knowledge of the ontological structure or search parameters (Collao, Diaz-Kommonen et al. 2003).

SOM analysis supports three modes of use: clustering, visualisation and a hybrid of both approaches. The process involves mapping data patterns to an n-dimensional grid of neurons. The grid develops an output space corresponding to the original input space as the process observes the original topology. Within this context, the SOM is used to output nodes (atomic information units) into topological representations of the original data input; these topological representations resemble soft-edged classes of nodes and are created through the SOM's self-organisation process (Collao, Diaz-Kommonen et al. 2003).

Although the discussion has thus far concentrated on ways to conceptualise knowledge, narrative still remains as one of the most primitive and fundamental ways in which humans construct, share and represent real-world knowledge. It is within this context that the following discussion explores narrative from the perspectives of learning (section 2.4.1) and new technology (section 2.4.2).

2.4. Narrative

David Lodge, in his essay *Narration and Words,* describes narrative as '*one of the fundamental sense-making operations of the mind and would appear to be both* of *g* groups so that members of the same group are more alike than members of different groups (Flexer A. 1999). This technique is widely deployed in areas that deal with large amounts of data as a method of automatically producing similarity clusters (SC). SCs may be comparable to terms in a thesaurus or concepts from an ontology; in that entities who display semantic similarities are closely related. In this way, SCs can be thought of as a visual representation of the objects and their relationships. However, an important difference is the rigidity exhibited by concepts and terms compared to the soft-edged clusters formed on a SC map. The clustering process develops a semantic space similar to a thesaurus yet the technique is automatic and supports a more ambiguous classification methodology.

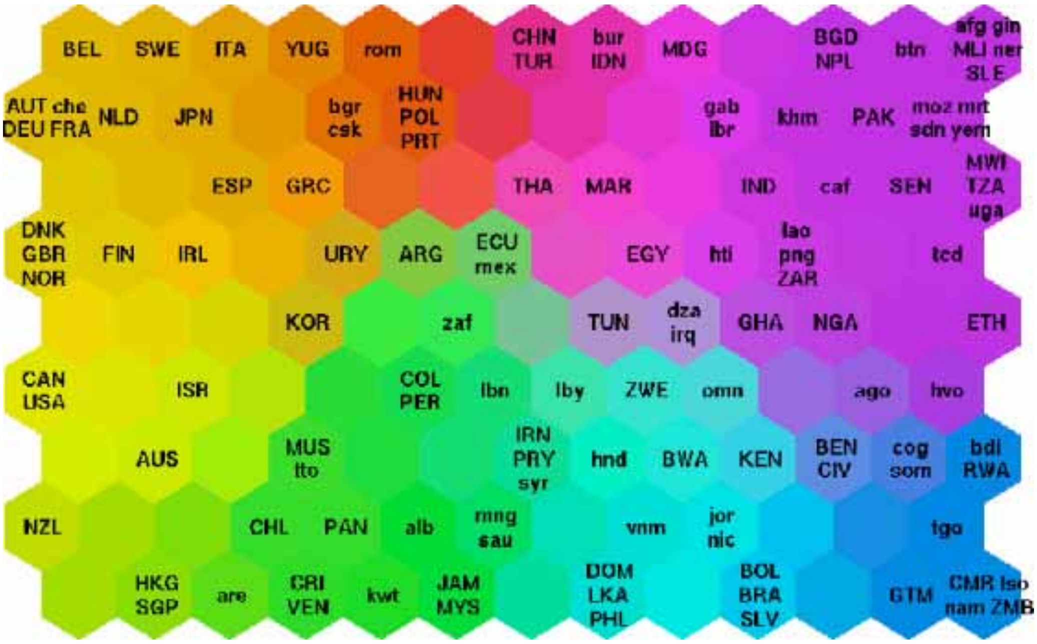


Figure 6: Shows a SOM visualisation of the world poverty map created by the Neural Networks Research Centre at Helsinki University (http://www.cis.hut.fi/research/som-research/worldmap.html). As a contrast, the more affulent nations are depicted in the tan-coloured clusters, while the less fortunate are illustrated in purple and blue. The *ambient* nature of a SOM is illustrated by the colour grading from one country to country.

be of one type or another and can exist amongst the soft-edged classes of neighbouring entities.

Artificial Neural Networks (ANNs), for example, are parallel computational models. The paradigm was originally inspired by the bioelectrical networks of neurons and synapses found in the human brain. In this way, ANNs aim to simulate the information processing capabilities of their biological counterparts. There are two types of ANNs, those that require human guidance and those that do not (Collao, Diaz-Kommonen et al. 2003). Supervised ANNs involve the extraction of a desired output result based on each input. This target result is used to guide the formation of new neural parameters and so train the network to learn the process currently under study. Supervised ANNs have proven effective in decision-making, object-recognition and forecasting. Unsupervised ANNs, however, require no guidance and the learning process is entirely data-driven. Unsupervised ANNs are largely used as techniques for classifying, organising, dimensionality reduction, sampling, vector quantisation, clustering and visualising large data sets. One form of unsupervised learning that was developed during the 1980s is called self-organising maps. The technique, which is discussed next, supports similarity clustering across large datasets.

2.3.1. Self Organising Maps (SOM)

The SOM algorithm was first described by the Finnish professor Teuvo Kohonen (Kohonen T. 1982) as a means of automatically arranging high-dimensional statistical data so that similar inputs are mapped according to their underlying semantics. SOM provides two types of cognitive functionality; the first is clustering and the second visualisation. Clustering is a method used to divide a set of n observations into a set

2.3. Semantic Clustering

Disciplines, such as information extraction, are supporting new approaches to ontology acquisition. Onto-extract (Sure, Akkermans et al. 2003), for example, supports semi-automatic ontology development through analysis of a large corpus of text. Computational approaches remove the burden of manual development while offering an interpretation of the underlying content free from human intervention. Further is the branch of computing known as 'soft computing' that differs from conventional computing techniques in that it proposes tolerance to imprecision, uncertainty and approximation, and which models its behaviour on the human mind. The principles of soft computing stem from Zadeh's 1965 paper that suggests fuzzy sets as a representation scheme and calculus for uncertain or vague understanding. Soft computing introduce the concept of a soft-edged or blunted boundary that marks a qualitative approach to categorisation and deviates from such categorical representations as assumed by conventions of philosophy and psychology (Díaz-Kommonen and Kaipainen 2002). The following describes some techniques, which fall under the heading of soft computing.

Automatically computed semantic models are void of human interpretation, yet support human understanding of high-dimensional data. The advantages of such techniques are evident when dealing with large data sets. Humans lack the ability to comprehend and extract meaning from very large bodies of data. This has led to much research having been conducted in trying to support human understanding of multi-dimensional data through semantic representation and increasingly sophisticated visualisation technologies. It is in this way that many techniques support a more ambiguous interpretation of artefacts, as such items are not required to

Semantic Indexing (LSI) or Analysis (LSA), for example, is one such method that was developed during the 1980's to improve detection of relevant documents on the basis of their underlying semantics. LSI, as described in (Deerwester, Dumais et al. 1988) and (Landauer, Foltz et al. 1998) is a method of statistically developing a semantic structure representing a corpus of text. It was developed to try and combat the inherent problems associated with syntactical search, chiefly ambiguity through polysemy[28] and synonymy[29]. Although term expansion and thesauri have helped tackle the problems associated with synonymy, polysemy has been more difficult to approach as it is inherent in natural language. One tool employed by information scientists is to use a Controlled Vocabulary to help disambiguate query terms. However, LSI utilises implicit higher-order structure of the association of terms in creating a multi-dimensional semantic structure of information. This structure can be used to transverse the data set according to the texts underlying semantics; furthermore the method favours queries based on a text's conceptual structure as opposed to specific term matching. The approach proposes that there is an underlying latent semantic structure in text documents which is obscured by the randomness of word choice with respect to retrieval. The method uses statistical analysis in estimating this latent semantic structure. Additional to LSI, the following section discusses some further approaches to automatically structuring content without the need for human intervention.

[28] Polysemy signifies a word or phrase with multiple meanings.

[29] Synonymy represents the semantic relation between two words or phrases that express the same meaning.

In contrast, Srinivasan, during his work on the *Village Voice* project, approached ontology development in terms of community participation and mutable knowledge structures (Srinivasan 2003). He referred to the concept as fluid ontologies, or 'flexible knowledge structures that evolve and adapt to communities' interest' (Srinivasan and Huang 2005). It is important to note two significant points about Srinivasan's work. Firstly, the community exhibited a low-level of computer expertise, i.e. it was a non-technical community. Secondly, fluid ontologies are, in all reality, informal hierarchically organised terms, similar to topic maps and reasoning engines are unable to interpret or sufficiently process knowledge structured in this way. However, he approached ontology acquisition from the point of community participation and in that succeeded in developing a flexible approach to structure collective knowledge.

Building ontologies from scratch is always an option open to the information designer. Naturally, this can be a barrier to approaching formal ontologies, because ontology development is a difficult undertaking and the process may become lengthy as a consequence. There are, nevertheless, several studies outlining development methodologies that have proved successful in the field. Some notable studies include (Cristani and Cuel 2004), (Fernandez, Gomez-Perez et al. 1997) and (Corcho, Fernández-López et al. 2003).

There is also a push to use software-generated ontologies to remove or shorten the time associated with ontology development. Often described as semi-automatic ontology development, this approach helps the ontology developer to circumvent the early phase of knowledge acquisition (suggesting initial domain concepts, etc.), which may have been carried out with a survey or with round the table discussions, and therefore offers a foundation on which to develop and organise knowledge. Latent by Gruber, are not created to fulfil a specific task or work practice. The difference between the representations, that of the data model and that of the ontology, emerges in the modelling process. When dealing with an ontology, for example, large amounts of effort and attention must be paid to the philosophical notion of a concept. On the other hand, a class or construct in a data model might be changed regularly to accommodate the development cycle.

Much research has been conducted in trying to develop new and innovative practices in knowledge engineering and conceptual modelling with the aim of reducing the complexity and overhead of ontology development. One such approach is the acquisition and reuse of existing ontology structures (Noy and McGuinness 2002). Assuming large ontology libraries exist, an ontology for a given application can be assembled from existing ontologies held within a library (Farquhar, Fikes et al. 1996). This technique helps to standardise a domain and inculcate common meaning within a particular discipline. Furthermore, it supports a Darwinian approach to the application of ontologies in that the best and most applicable ontology will survive for any given instance. Currently, there are a number of ontology libraries available on the Web; examples include the Ontolingua[26] and the DAML[27] ontology library. As ontologies move from academic circles to industry, increasing numbers of commercial ontologies are becoming publicly available, e.g. UNSPSC and RosettaNet (Noy and McGuinness 2002).

[26] The Ontolingua library can be found on Stanford's site at http://www.ksl.stanford.edu/software/ontolingua/.

[27] The DAML Ontology library is available at http://www.daml.org/ontologies/.

evaluate the National Cancer Institute (NCI) medical thesaurus in light of its conformity to both terminological and ontological principles. Their work suggested that although there are a number of irregularities, the NCI thesaurus provides an excellent resource when used internally, i.e. when used in a close system. If however the thesaurus was to be applied in other contexts, the authors' suggested that '*a considerable effort will have to be made in order to clean up its hierarchies and to correct the definitions and ambiguous terms*'.

The following will discuss ontology acquisition[25] - ways in which ontologies are created, developed, adopted, refined and finally employed.

2.2.3. Ontology Acquisition

Unlike a data model where the representing structure generally functions in a 'single' specific application (i.e. database or banking system), an ontology, as discussed in 2.2.2, is supposed to be considered from a more universal perspective. If approached correctly, a domain should be rendered separately from the problems and tasks which characterise the domain. On the contrary, data models, when once instantiated, often operate within a single architecture; they are not usually developed to be shared by other applications, and therefore encapsulate domain, tasks and problems within a single application architecture. The semantics of a data model, as described by Spyns et al., often comprise of an informal agreement between developers; frequently amendments occur when warranted and without large amounts of consideration (Spyns, Meersman et al. 2002). By contrast, ontologies, as suggested

[25] The title, ontology acquisition, was chosen because the discussion does not focus on any one approach to acquiring or indeed developing ontologies. Rather the intention here is to introduce some way to approach the creation, adoption and use of ontologies.

developers with the ability to specify first order logic constraints, i.e. properties beginning with value restrictions[19] and moving onto more expressive object relationships, such as symmetric[20] properties, transitive[21] properties, inverse[22] properties, functional[23] properties, etc. (McGuinness 2002). It is up to the ontology developer on how to express a concept. Using more detailed relationships, however, help to build intelligence into an ontology, enhancing the ability of software to reason across the knowledge base[24].

Complexity, as will be argued in (see section 6.3.3), comes at a price. It is difficult to develop highly structured ontologies, comprised of unambiguous concepts and multiple logical properties. Systems may operate reasonably when used in a specific instance (or closed system) but when introduced as part of a larger ecology (as is the intention with the Semantic Web), concepts must be defined correctly; otherwise each additional system will reason across incorrect or flawed knowledge. This was shown in (Ceusters, Smith et al. 2005) where the authors attempted to

[19] A simple example of a value restriction is the wheels property mentioned previously being only able to take a number between 2 and 10, 2 being a motorbike and 10 a large truck.

[20] Symmetric properties states that if A is related to B through P then B is equally related to A thought P (Horridge, Knublauch et al. 2004).

[21] Transitive properties are based on the simple logic principle (or syllogism): Humans are mortal. Greeks are human. Therefore, Greeks are mortal (Shirky 2005).

[22] Each ontology property (that takes an object and not a literal) can have an inverse. For example, the property hasChild has an inverse property of hasParent. If Father hasChild then Child hasFather (Horridge, Knublauch et al. 2004).

[23] A functional property is a unique relationship between two concepts. For example, one child can have only one birth-mother, therefore "child hasBirthMother" mother is a functional property (Horridge, Knublauch et al. 2004).

[24] See (Horridge, Knublauch et al. 2004) for a list of properties available to the ontology developer when making use of the OWL ontology language. It is not the intention of the author to describe the purpose or application of formal ontologies but rather to illustrate some of the approaches open to the information designer when wishing to capture and organise knowledge. There are, however, some excellent studies on the state of the art of both ontology languages and the accompanied engineering methodologies; most notable is Corcho's, Frenandez-Lopez and Gomez-Perez (Corcho, Fernández-López et al. 2003).

although the relationship between 'fort and castle' may be understood, implicitly, by people, computers are unable to process information encoded in this way. This is because computers lack the stacks of general knowledge that humans naturally develop throughout their lifetime (see (McCarthy 1984))).

However, it is up to the developer of the taxonomy to decide whether or not to include a formal hierarchy. The next point on the spectrum deals with strict subclass hierarchies. As with Object Orientated Programming (OOP), strict formal hierarchies are necessary for exploitation of inheritance (McGuinness 2002). Strict subclass hierarchies adheres to the rule that if A is a subclass of B, and if an object C is an instance of A it stand that C is also an instance of B. A simple programming example is that of a car being a subclass of a vehicle and a Ford car being an instance of both. There next point introduces the concept of Frames[17], an approach to knowledge modelling developed by Marvin Minsky. Frames lie outside the scope of this thesis, but are comparable, in many ways, to classes in OOP and concepts in an ontology.

Moving further from left to right, McGuinness states that '*as ontologies need to express more information, their expressive requirements grow.*' Some ontology languages[18] allow developers to state arbitrary logical statements while others provide

[17] Frames, developed by Marvin Minsky, were an attempt to clarify the semantics of nodes and links (Graph Theory) by using a method for representing real-world knowledge. He conceived that conceptual encoding in the human brain is not concerned with strictly defining concepts but in finding examples of categories that concepts fit into. He described frames as 'data structures for representing stereotyped situations' (Minsky 1975). A frame could therefore be compared to a prototypical object or prototype. Like classes in OOP, for example, frames include property information, i.e. The Vehicle class presented earlier may include the properties wheels, and an instance of that class could be type Ford and have four wheels.

[18] It is important that an ontology developer identifies all specific requirements of an ontology. This is because ontology languages differ greatly, particularly when it comes to expressing concepts. For example, RDF, although ideal for describing web resources via graph based logic models, is not as powerful as OWL, which comes in three flavours lite, DL and Full. The expectations of an ontology will dictate the language used to define it.

readable format, usually First Order logic (Gamper, Nejdl et al. 1999). Thirdly, an ontology states knowledge explicitly, meaning that the types of concepts used, whether explicit or tacit in a real-world sense, should be explicitly defined. Finally, an ontology must represent a shared understanding of a domain; the knowledge which an ontology embodies should be shared by a group, and not restricted to a single individual (Fensel 2000). In summary, a formal ontology explicitly specifies some agreed upon conceptualisation in a formal, i.e. machine readable, language.

The basic hierarchical structure of an ontology is taxonomic, (splitting Figure 4) a recognised and popular way to structure digital content. The mathematical basis is that of a tree, beginning with a root node and structuring out into relevant branches. Each branch is a subset of the previous node, but represents a more specific term. There are only three principle concepts: Node, Parent and Child. A node represents a subject, whether conceptual or physical, a Parent represents a broader term and a Child signifies a more specific one. The tree structure exposes similarities among groups of nodes, their parents and children. The nodes are grouped according to likeness in physical characteristic, role, function, structure, etc. Taxonomies are essentially subject-based hierarchical classification structures and differ only from controlled vocabularies by a relationship formation. Due to straightforwardness and ease-of-use, taxonomic classification is a widely used organisational structure by information architects and is visible throughout the web (Google and Dmoz) (Rosenfeld and Morville 1998). Many early web specifications of term hierarchies such as Yahoo, however, did not display a strict and formal 'is-a' hierarchy. McGuinness distinguishes this point on the spectrum as it seems to capture naturally occurring taxonomies on the web. As with thesauri, without an explicit or true 'is-a' hierarchy, certain deductive uses of ontologies become problematic. For example, Because framework terms are non-index-terms,[14] they may present problems to software agents relying on an exact "is-a" hierarchy.

Although thesauri exhibit weaker semantics they are still widely used across a range of domains to help structure and organise content. In the heritage domain, for example, thesauri are consistently being developed to help with the problem of classification. The Getty Institute's vocabulary program[15] maintains several thesauri relating to the classification of artefacts. Similarly, English Heritage[16] also developed a series of thesauri in support of major computerisation projects. The goal of much of this work is to provide an organising principle when structuring digital content.

There is much research and investment being placed in creating new approaches and techniques for managing and interacting with digital content. Although simple, ontologies can help with the identification and retrieval of digital resources, more detailed specifications, such as structured ontologies as will be discussed next, have emerged as a powerful alternative to defining digital content.

2.2.2. Structured Ontologies: Adding the ability to reason

Dieter Fensel describes four points he uses to identify or define formal ontologies. Firstly, a conceptualisation refers to an ontology's ability to reflect an abstract model of some real world phenomenon, defining all relevant concepts and relationships of the phenomenon. Secondly, an ontology must be represented formally in a machine

[14] Non-index-terms are thesauri terms that allow groupings of terms according to some subject or function. However, as the name suggests non-index-terms should not be indexed or used for searching.

[15] For a list of thesauri see http://www.getty.edu/research/conducting_research/vocabularies/

[16] A complete list of English Heritage thesauri may be found at http://www.englishheritage.org.uk/thesaurus/frequentuser.htm.

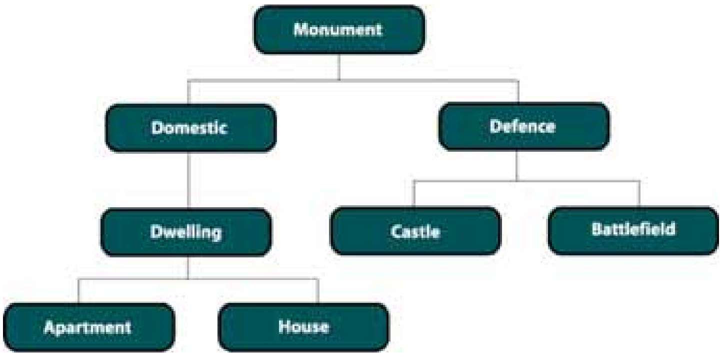


Figure 5: A section of the English Heritage monuments thesaurus[13].

Thesauri, Figure 5, are used to describe things and concepts consistently. The most famous example is, perhaps, that of *Roget's Thesaurus*, compiled in the early nineteenth Century. A thesaurus groups related terms together and provides cross-referencing to other groups of terms that may be relevant to the subject. A preferred term is supplied to avoid ambiguity and, by arranging terms in a hierarchy, the selection of more general or specific terms is provided for. The proliferation of this type of conceptual model can be observed through the development of two ISO standards offering guidelines in the creation of both monolingual (ISO2788 1986) and multilingual thesauri (ISO5964 1985). However, although a degree of informal specification can be established through the use of narrower and broader terms, typically thesauri do not provide an explicit hierarchy. McGuinness maintains that without true subclass relationships, certain kinds of deductive uses of ontologies become problematic. Within this context thesauri do not necessarily exhibit a true "is-a" hierarchy (such as a Tom is-a human) rather the hierarchy will contain framework terms that are artificial expressions to allow objects be grouped together by their uses.

[13] This diagram was compiled from the English Heritage Monument's Thesaurus available at http://thesaurus.english-heritage.org.uk/thesaurus.asp?thes_no=1.

and value restrictions. There is, however, a distinct difference between simple and structured ontologies, as illustrated by the darker shading in Figure 4. This difference is based on stronger semantics; or the ability of reasoning software or inference engine to reason across a specific ontology. Simple ontologies, or metadata models, are generally developed for human consumption, while formal ontologies are designed for both human and machine consumption, i.e. they are developed by humans with the intention that machines can digest and interpret their meaning. Using McGuinness' spectrum as reference, the following section will discuss several approaches to the creation and use ontologies when structuring and representing real world knowledge.

2.2.1. Simple Ontologies: Structuring digital content

Simple (informal) ontologies, sometimes referred to as terminological ontologies (Gamper, Nejdl et al. 1999), use natural language to define terms. Artificial agents – reasoning engines or software systems that exhibit some form of autonomous agency - are unable to interpret and sufficiently process natural language statements. Therefore, simple ontologies are used predominantly in human centred platforms, often to organise information, provide site support or narrow search criteria. The simplest notion of an ontology is that of a controlled vocabulary (CV); a closed list of terms used to characterise a domain, e.g. a library catalogue (McGuinness 2002). Similarly a glossary uses natural language statements to describe a list of terms particular to a specific domain.

description by asserting that '*an ontology necessarily entails or embodies some sort of worldview with respect to a given domain. The worldview is often conceived as a set of concepts, their definitions, and their inter-relationships; this is referred to as conceptualisation*' (Uschold and Gruninger 1996). Both definitions are consistent with the usage of ontology as a set of set-of-concept-descriptions, but it is used in a different sense of the meaning in philosophy (Gruber 1993).

In trying to clear up some of the ambiguities of the term ontology, McGuinness designed an ontology spectrum (which will be cited throughout this thesis), where several ontologies are viewed according to the detail of their specification (McGuinness 2002).

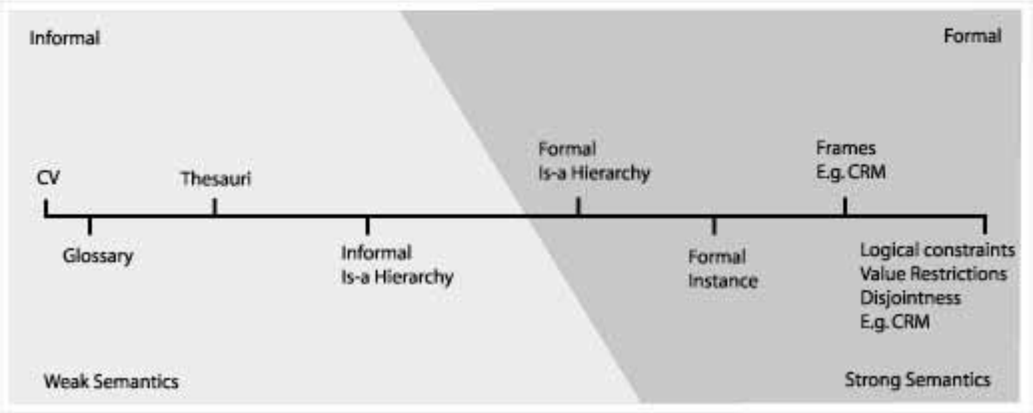


Figure 4: McGuinness' ontology Spectrum (McGuinness 2002).

In Figure 4, she divides the spectrum between simple, term-based ontologies and formal, detailed ontologies. This division is based upon the ability of an ontology to clearly and unambiguously identify concepts. The spectrum classifies the simplest notion of an ontology as a catalogue, providing some conceptualisation and interpretation of terms. The detail increases towards the right hand side of the spectrum; thesauri offer synonym relationships, taxonomies provide generalisation and specialisation, while more logical models exhibit property types, formal instances

The more descriptive the metadata model, it could be argued, the more advanced the level of understanding, and it is within this context that recent years have seen a rebirth of interest in the use of AI. Many of the approaches that were cultivated in labs during the heyday of AI are being re-examined as a possible answer to the problems facing contemporary information design. Ontologies, for example, once suggested as a way to encode and re-use an expert's knowledge, are being proposed as the backbone to the semantic web (McGuinness 2002).

2.2. Ontologies

The term *ontology* first arose in philosophy (Gruber 1993) and, for the sake of simplicity, may be defined as '*the study of the kinds of things that exist*' (McCarthy 2007). It is within this context that ontologies '*carve the world at its joints'* (Chandrasekaran, Josephson et al. 1999). However, in recent years, interest in ontologies has spread from disciplines surrounding philosophy to more computational areas[12] such as knowledge management, information science, qualitative modelling and intelligent systems integration. The basis of this interest, amongst others, lies with an ontology's ability to model a shared understanding. However, the term 'ontology' suffers from ambiguity and seems to generate a lot of controversy, particularly within the AI and KM sectors.

Gruber's widely cited definition simply states, '*an Ontology is a formal, explicit specification of a shared conceptualisation*' (Gruber 1993). Uschold & Gruninger give a more detailed definition stating that, '*Ontology is a term used to refer to the shared understanding of some domain of interest*'. They expand on Gruber's

[12] See (Guarino 1998) for comprehensive listing of fields embracing the use of ontologies.

hardware and software to a more fluid and social content-driven movement (Wolfe 2000). This shift focuses on cultural dynamics and co-ordination and places metadata squarely in the sphere of communications and knowledge management practices. The use of metadata adds a human layer of intervention and interpretation into information systems, as resources are assessed according to the way they are represented, and interpreted by the metadata that describes them.

Friesen contends that the shift in emphasis implied by the application of metadata can be understood as a shift from data processing to the creation and interpretation of information or knowledge. From this perspective, data, information and knowledge are viewed as a hierarchy with each layer being differentiated from the last through a process of interpretation and arbitration leading to increased human understanding, intention and purpose (Friesen 2002). Although metadata can provide a greater context for an information resource, it seems misleading to say that a singular piece of metadata characterises an interpretation of meaning or purpose of a digital resource (Friesen 2002). To have relevance, raw data needs to be transformed or interpreted into information or knowledge, as John Sowa states '*meaningless data cannot acquire meaning by being tagged with meaningless metadata*[11]' (Sowa 2000). Nevertheless, without a clear indication and understanding of how metadata are to function, an information resource described by metadata could be regarded as of little use. However, with the development of metadata models, descriptive vocabularies, etc., resources or data can be supplied with greater interpretation, context and purpose.

[11] It could be argued that with the rise of social tagging, and the creation of community folksonomies (Mathes 2004), data can indeed acquire meaning by being tagged with meaningless data. However, for social tagging to be effective, a broad community of users is required, and the data cannot, as yet, be structured in a formal machine-readable way. The use of tagging in this way furthers the ability of a group to decide on the meaning of a resource.

2.1. The duality of Knowledge

It is often argued, when codifying knowledge for reuse, that knowledge may be thought of as a duality, consisting of two composite parts, explicit knowledge, that which may be stated or articulated, and tacit knowledge, that which the knower can do but may be unable to express in words (Hildreth and Kimble 2002). Implicit knowledge is the knowledge that is most difficult to define, often tacitly performed, and identified by the Greeks as *techne*, know-how as opposed to *episteme*, know-what (Taleb 2007). Although, in principle, explicit knowledge may be captured, codified and shared digitally, the elusive nature of tacit knowledge means it is, at best, extremely difficult to capture and share and, at worst, impossible. This is an issue that the knowledge management community is constantly grappling with, and has led to both technology and people-centric approaches to organizational design (Wolfe 2000).

The technology approach deals with explicit knowledge and emphasises the design and deployment of computer based systems to affect collaborative workspaces. The people-centric approach takes a cultural viewpoint and aims to encourage and capture the interaction of tacit knowledge, and the overall stocks of tacit knowledge that social agents have to bring to bear when performing tasks (Wolfe 2000). The two approaches, however, often overlap to some degree and this has led to a shift in information management thinking from an engineering-based approach to one which emphasises information as the product of social actors embedded within cultural networks. Wolfe maintains that the rigidity of earlier computer systems often alienated the user and that through better information management and the use of metadata there is a valid and complementary shift from a world dominated by

2. Knowledge Representation

This chapter examines knowledge representation, as informal metadata descriptions, formal machine-readable ontologies and automatic semantic clusters. The purpose of the chapter, however, is not to suggest one approach to knowledge representation over another, but rather to provide the reader with an overview of an interesting and dynamic area of research. Knowledge can be presented as a duality, consisting of both the tacit and explicit forms of knowledge (as discussed in section 2.1). Continually, practitioners of knowledge management are coming to terms with how to capture and codify tacit knowledge. It is within this context that the author introduces several approaches to knowledge representation, considered under the heading of ontologies (section 2.2), as valid options open to the information designer when wishing to structure real-world knowledge. Many of these approaches are being heralded as the panacea or 'silver bullet' to both the problems of knowledge management (KM) and the semantic web (Fensel 2000). The chapter will then go on to examine a machine learning approach to semantic clustering (section 2.3). This, again, is comparative to manual approaches within the discipline of ontology engineering but can reduce the time and effort required of the developer. The chapter concludes with a discussion on narrative (section 2.4). The author argues that narrative is still the main way in which humans structure, communicate, share and represent knowledge. Narrative is therefore examined as a form of knowledge representation and discussed from the perspective of learning (section 2.4.1) and new technology (section 2.4.2).

services, such as automated reasoning. Next narrative was introduced, and online communities were considered as learning communities, providing members with opportunities beyond communication.

The next chapter will discuss knowledge representation from the perspective of tacit and explicit knowledge, and introduce several approaches to organising knowledge for retrieval and reuse. The chapter concludes with a discussion on narrative, one of the most primary means by which people construct and share knowledge.

In chapter 6 the author investigates the process of collaboration by looking at two approaches to structuring community-based knowledge. The chapter investigates RQ2, comparing the approach of researchers at KMI, whilst working on the Bletchley Park forum, with that of researchers at DIT, when developing the Explorer forum. The chapter aims to identify the important and relevant aspects of community involvement by exploring both approaches under the headings, population, community models and collective knowledge vs. formal ontologies.

In answering RQ3, chapter 7 approaches the use of structured ontologies versus that of more ad-hoc approaches to data modelling. Within this context, the chapter focuses on the representation of narrative, examining two different approaches to structuring digital narrative for use in a community based environment. The chapter goes on to compare approaches under the headings, reusable narrative structures, consistency checking and interoperability, and concludes with a brief discussion on the use of ontologies when compared with an implementation of a less formal method to organising user-generated content.

The final chapter presents conclusions and introduces some interesting areas, which warrant further research.

1.6. Summary

This chapter introduced the background to this thesis. Knowledge representation has become an area of much contemporary interest as researchers and practitioners move to handle information in increasingly innovative ways. Web, 2.0, or the social web, has preceded the Semantic Web, where communities use approaches such as social tagging to organise community-based content. At the same time there is much research being conducted in developing ontologies to support more advanced

The chapter concludes with a discussion on narrative, one of the fundamental ways in which human's develop and exchange real world knowledge. The discussion focuses on narrative from two different perspectives, learning and new technology. The first aims to emphasise the importance of narrative in learning, as a structural framework for knowledge and as a vehicle for knowledge transfer. It discusses narrative from a pedagogical standpoint, and highlights the importance of narrative in socially constructing knowledge. The second, new technology, deals with writing narrative for a hypertext platform and discusses the aesthetics of the medium.

Chapter 3 focuses on the literature surrounding online communities. The chapter illustrates that the prominence of the online community has led to increasing diversity in type and use. Most notable are the differences between work-enabled communities, sustained through goal-orientated work practices, and communities developed through a shared interest. The chapter concludes with a discussion on some of the more traditional approaches to online community design.

Chapter 4 presents a discussion on several of the topics raised in the previous two review chapters. The discussion attempts to bridge the disciplines of knowledge representation and online communities. Several questions are developed to help structure the remainder of this thesis. Each question is addressed separately in chapters 5, 6 and 7.

Chapter 5 compares the soft-ontology platform, an approach developed by researchers at UIAH, with that of a standard thesaurus implementation, as carried out by researchers at DIT. The chapter seeks to answer RQ1, discussing each approach under the headings, collective classification, community models, method of contribution and, finally, limitations.

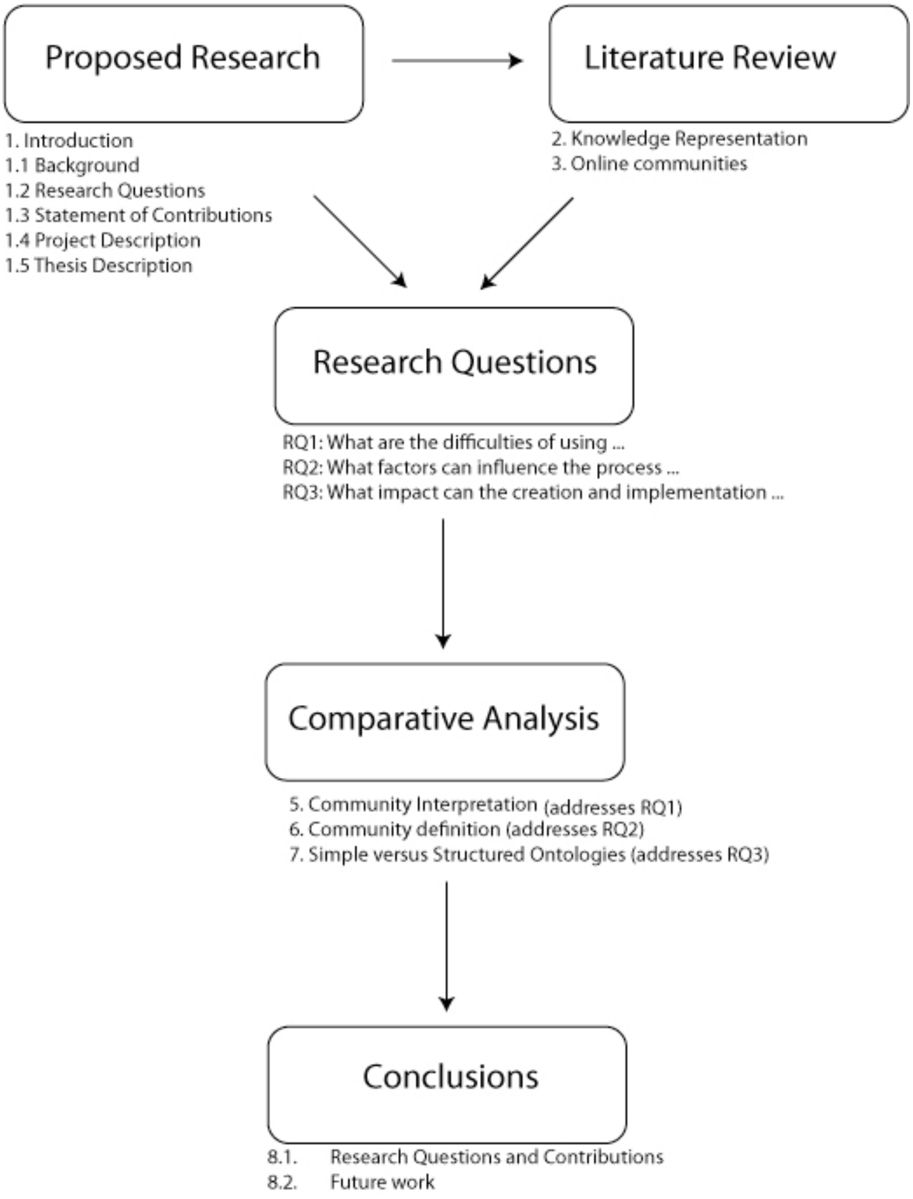

Proposed Research
Literature Review
1. Introduction
1.1 Background
1.2 Research Questions
1.3 Statement of Contributions
1.4 Project Description
1.5 Thesis Description
2. Knowledge Representation
3. Online communities
Research Questions
RQ1: What are the difficulties of using ...
RQ2: What factors can influence the process ...
RQ3: What impact can the creation and implementation ...
Comparative Analysis
5. Community Interpretation (addresses RQ1)
6. Community definition (addresses RQ2)
7. Simple versus Structured Ontologies (addresses RQ3)
Conclusions
8.1. Research Questions and Contributions
8.2. Future work

Figure 3: Structure of this Study.

Chapter 2, knowledge representation, provides an overview of several approaches to organising knowledge and structuring community based content. The chapter opens with a discussion on explicit and tacit knowledge and goes on to discuss the differences between simple and more structured ontologies. The chapter then provides an overview of some techniques used to automatically create semantic models based on a branch of computing known as 'soft computing'.

and *community definition* inform the chapter *simple versus structured ontologies*, which explores the benefits of each method when approached from a community perspective.

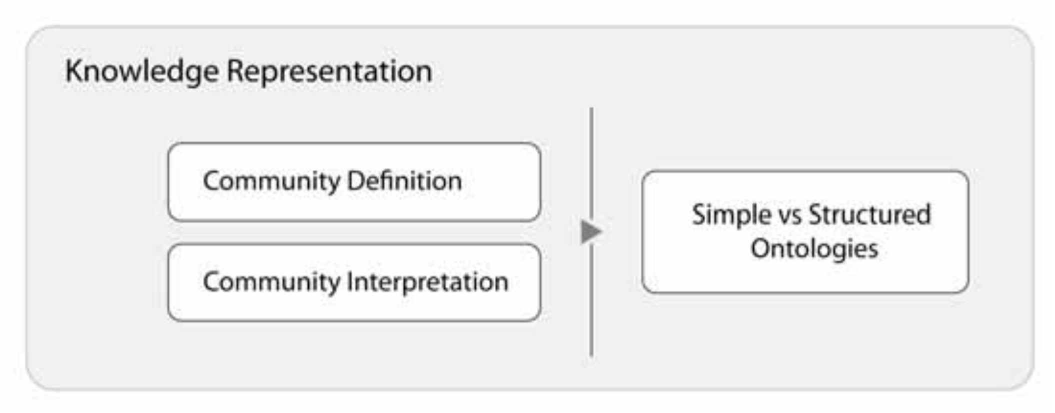


Figure 2: A Conceptual overview of this thesis

Figure 3 illustrates the structure of this study. The current chapter introduces some of the broader issues involving knowledge representation. This chapter presents the research questions, contributions and includes a description of the project that served as a platform for this research.

narrative content. While the methodological goal aimed to investigate and subsequently develop appropriate methodologies for supporting the community's activity when participating in a CIPHER forum.

In the context of this thesis, each partner chose a different approach to realising the application goal of the project. This resulted in several different forums being created with a range of different technologies. Nevertheless, all of the forums utilised some form of knowledge representation to either structure newly created content or present the user with a new way of accessing existing content. Furthermore, the methodological goal of the project was realised through the creation of processes to support the activity of each community. This approach did not specify any single technology but rather offered a range of technologies for creating sustainable online communities. All the examples put forward by the author in this thesis are therefore specifically related to the culture heritage domain. However, this is not to say that the approaches developed in line with the project can only be applied to the domain of cultural heritage. There are aspects of cultural heritage, the importance of time and events for example, that must be considered by the information designer. Nevertheless, the overarching approaches to both community based platforms and the use of narrative in a community environment are domain independent and can be adopted for other domains of interest.

1.5. Thesis Description

Figure 2 outlines the conceptual overview of this thesis. The main research theme is knowledge representation when applied to online communities. Therefore, the community and the impact of the community on the approach to and outcome of knowledge representation are examined. Both chapter's *community interpretation* represented on the Carta Marina. See (Díaz-Kommonen and Kaipainen 2002; Díaz-Kommonen and Kaipainen 2002; Collao, Diaz-Kommonen et al. 2003) for more information on the Carta Marina forum.

1.4.3. Technology Innovation in South Central England

The Bletchley Park CH forum was developed by researchers at the Knowledge Media Institute (KMI), England and aimed to enable a group of tour guides to research and investigate the activities of the code breaking facility of Station X at Bletchley Park. The approach involved the creation of an ontology of narrative to structure narrative content, and the development of a domain ontology to reflect the six year period of Station X in the Second World War. Both ontologies were used to organise stories of events and activities that occurred during the time of station X. For more information about Story Fountain and the Bletchley Park Forum see (Mulholland and Collins 2002; Mulholland, Zdrahal et al. 2002; Mulholland and Zdrahal 2003; Mulholland, Collins et al. 2004)

1.4.4. Shared Heritage of Central Europe

This Central European CH Forum was developed by the Czech Technical University (CTU), Prague in collaboration with GIS, Austria and provides online access to a large volume of data concerning historical sites in the Czech Republic and Austria. Several technologies were utilised in developing the forum. The use of narrative, and ontologies to identify and define narrative content, helped authors to structure and organise their CH stories.

It was within the framework of several active forums that both the technical and methodological goals of the project were investigated. The technical goal involved developing innovative approaches to exploring existing content and creating new

1.4.1. Irish Cultural and Natural Heritage

The Explorer cultural heritage (CH) forum was developed by researchers at DIT with the aim of providing a mechanism for the public to record cultural heritage stories. Stories in this sense involved personal and sometimes collaborative accounts[9] of Irish history and prehistory. The aim of the approach was to reflect earlier efforts at hypertext fiction whereby the reader explores an unfolding domain through a series of dynamic narratives. For more information see (Kilfeather, McAuley et al. 2003; Kilfeather, McAuley et al. 2003; McAuley and Kilfeather 2005). Further, Appendix B: Developing Explorer provides an overview of the technologies, approaches and techniques used to develop the Explorer forum.

1.4.2. Nordic Heritage, Storytelling and Historical Artefacts

The Nordic CH forum was developed by researchers from UIAH, Helsinki Finland and drew inspiration from two celebrated cultural heritage artefacts, the Carta Marina[10] of 1539 and 'A description of the Northern Peoples', 1555. Olaus Magnus, the last Catholic bishop of Uppsala, Sweden, created both artefacts before his death in 1557. The Carta Marina is acknowledged as the first comprehensive description of the landscape and people of the Nordic region. Additionally, the map displays a host of monstrous mythical figures inhabiting Nordic regions. It is generally considered that his 'A Description of the Northern Peoples' ("Historia de gentibus septentrionalibus"; Rome; 1555) is a commentary on the map. The forum used new technologies and innovative approaches to empower the user to explore the narrative

[9] A complete narrative presentation composed of stories sometimes written by several authors.

[10] A Carta Marina forum was developed around a digitized version of the Carta Marina map, which can be found at http://cipher.uiah.fi/forum/materials/carta_marina/annotations/

initiative to help communities create and explore regional cultural heritage, and on which the author collaborated as a software developer.

1.4. Project Description

The CIPHER project objectives, as outlined in the original proposal document, involved creating innovative methodologies and technologies to support the development and continued maintenance of self-sustaining cultural heritage forums (CIPHER partners 2001). A forum, in this context, is thought of as a virtual space in which communities are encouraged to explore, research and build content. The project comprised of six partners, who developed four online forums each representing a specific region of European cultural heritage. Each forum's aim was to support several different community models, with different population sizes, levels of technical expertise and experience of life within an online community. The project objectives were divided into more specific goals categorised as application, methodological and technical. The application goal of the project was represented by the creation of the four online CH forums as illustrated in Figure 1.

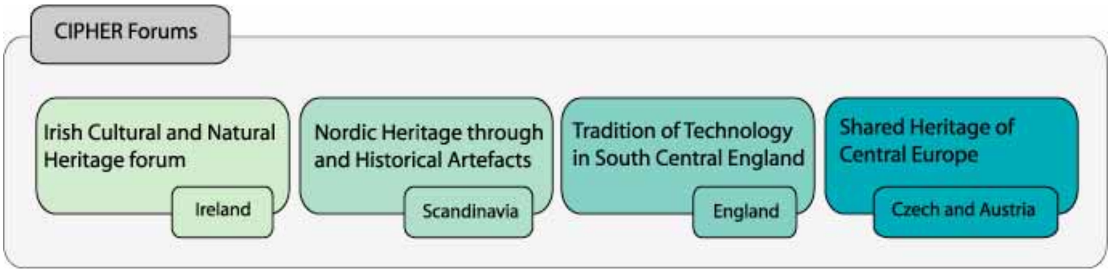

CIPHER Forums
Irish Cultural and Natural Heritage forum
Ireland
Nordic Heritage through and Historical Artefacts
Scandinavia
Tradition of Technology in South Central England
England
Shared Heritage of Central Europe
Czech and Austria

Figure 1: The four CIPHER Cultural Heritage forums.

The forums were:

RQ2: *What factors can influence the process of knowledge engineering when involving a community of non-technical users?*

RQ3: *What impact can the creation and implementation of a structured ontology have when representing narrative concepts?*

1.3. Statement of Contributions

Several contributions were identified as demonstrating some of the broader aspects of knowledge representation when applied to community-based environments.

Contribution 1: The first contribution presents a comparative analysis on traditional and more contemporary approaches to classification. The basis for this analysis is the development of computer-mediated communication and, as a consequence, the ability of the information designer to reach, and include, the community at every stage of the classification process.

Contribution 2: The second contribution presents an examination of how community factors can impact the approach to and the outcome of knowledge representation.

Contribution 3: The third contribution presents a consideration on the use of ontologies, both simple and more structured, as a platform to support the organisation of user-generated narrative content. In a broader context, the author suggests that many of the concepts narrative embodies are evident in other forms of user-generated content, such as blogs, wikis, which often tend to follow a narrative thread.

Each contribution is based upon several real examples that were established and advanced during the work on the CIPHER (Communities of Interest to Promote Heritage of European Regions) project, a two and half year European Commission and novelist David Lodge suggests) crucial to the way in which people construct inner representations of external experience. He describes it as one of the 'fundamental sense-making operations of the mind' (Lodge 1990) and when discussed from the point of view of learning is aligned with the constructivist perspective (Mulholland and Collins 2002). Through narrative people do not passively absorb abstract facts but instead construct meaning from knowledge and experience. The study of narrative is a broad area of research, much of which lies beyond the scope of this thesis. However, it is difficult to ignore the importance of narrative when exploring ways to represent collective knowledge. This is because written narrative, both as informal discourse and formal presentation, is still considered one of the main conduits for knowledge within community-based environments.

1.2. Research Questions

These are some of the broader issues that will be examined in this thesis. However, the author does not present this work as a solution to the problems of knowledge representation when approached with broad-based or diverse community models, but rather as a study of the possibilities open to those wishing to cultivate collective knowledge with successful approaches and more intelligent technology platforms.

The following research questions help to structure the discourse. All three are compiled in chapter 4, and each is individually addressed in chapters 5, 6 and 7. They are:

RQ1: *What are the difficulties of using traditional classification methodologies when approaching community-based platforms?*

and have traditionally provided a focal point for social groups to form, develop and preserve a sense of collective identity (Brown and Duguid 2002). They cite the creation of the 'British Royal Society' and the *zine* culture that emerged from the sixties as examples. More recently, however, the proliferation of web communities such as Flickr[7], YouTube[8] and, of course, Wikipedia has illustrated that communities are forming around user-generated content. These communities are publishing and sharing increasing amounts of digital content in a variety of media formats. It is within the context of community involvement that this thesis will explore knowledge representation, examining both the formal and less formal approaches to representing knowledge, as exhibited by simple and structured ontologies.

1.1.3. Narrative

This thesis will also examine narrative, as an organisational and educational tool. Stories are good at presenting things sequentially and they help to make diverse information coalesce into a structured account (Brown and Duguid 2002). The storytelling tradition has allowed communities to circulate information and individuals to draw on the collective knowledge or experience of the group. In technical communities, for example, people use storytelling as a way to present a coherent account of a problem (Brown and Duguid 2002). Likewise, solutions are discussed and presented in a similar narrative format. This informal knowledge exchange is central to the activity of an online community. It is (or as the academic

[7] Flikr, http://www.flickr.com/, is the popular photo-sharing website and a good example of contemporary approaches to online communities.

[8] The popular video sharing website, http://www.youtube.com/, is another example of community's developing around user-generated content. It could be argued that this is the true advantage of the web, as unlike broadcast media, the web, especially web 2.0, provides a platform for users to contribute content.

henceforth classified. They may decide to place several tags on a single resource but the relationship between those tags often remain (explicitly) undefined. Indeed the advent of technology and approaches, such as social tagging, support 'a new ease of assembly' (Shirky 2008). It is a bottom-up approach to information management performed exclusively by the online community.

In contrast, web taxonomies (Dmoz[5] aside) are generally developed from the top down. The same, it could be argued, is true of the more recent developments in ontology engineering, where the process requires substantial investment on behalf of a small, often esoteric, group of experts. Furthermore, unlike taxonomies, knowledge must be explicitly expressed by concepts, properties and their inter-relationship. The ontology must be tested, refined and evaluated before it can be put forward for use in a larger ecological instance. The 'web 2.0' paradigm, however, is a fundamental shift in this thinking, as knowledge which was once represented by specialised data handlers, is now thought of as a form of public mark-up. This shift in thinking demonstrates formerly discrete schools of thought converging on the subject of semantics.

While social tagging has demonstrated that users are willing to contribute to jointly organising content, the success of Wikipedia[6] has illustrated that people are also willing to contribute to a collective knowledge base. From this perspective, Brown and Duguid maintain that documents exhibit a 'community-forming character'

[5] The open directory project DMOZ, www.dmoz.org, is a collaborative effort to develop a community-based directory. The approach reflects many of the principles underlying Web 2.0 in that the community acts as editor, maintaining the directory as a resource for others.

[6] In a similar approach to both Flickr and Youtube, Jimmy Wales' Wikipedia, http://www.wikipedia.org/, which emerged from the editorial model of Nupedia (Shirky 2008), espouses user-created content but unlike either Flickr or Youtube, the model is developed upon knowledge contributed in essay format on a wide range of subjects.

push the semantic web into its second phase of application development (Frauenfelder 2004). While more recently Oracle's uptake of RDF and OWL indicates a shift taking place for the semantic web from academia into industry. However, there has been little by way of practical implementation, as yet, and for the most part, semantic web tools are not in everyday use.

1.1.2. Online Communities

There are those, Clay Shirky (Shirky 2005) for example, who suggest that ontologies are just not all that useful. In this context, 'web 2.0' (O'Reilly 2005), or the social web (Gruber 2007), has in many ways preceded the semantic web. This emergence is predicated on substantial community involvement and the collective organisation of web content. Digg[1], for example, is a popular collaborative news website where users suggest stories and other users rate those stories. Similarly, Delicious[2], the social book-marking site, pursues the concept of social tagging as a way for online communities to collectively organise web resources. Social tagging has further paved the way for the emergence of the folksomony (Voss 2007), or subject-based taxonomy, created by community activity. Unlike the semantic web, or indeed web taxonomies such as Yahoo[3] and Google[4], the use of social tagging removes the need for a predefined knowledge representation (Shirky 2005). Users simply tag resources with whatever labels they think appropriate and the resource is

[1] Digg, the popular news website, is driven by community contributions and can be found at http://digg.com/.

[2] The popular social bookmarking site Delicious, http://del.icio.us/, provides an easy way for users/community members to tag online resources with subject keywords for use by other users.

[3] The Yahoo Directory can be found at http://dir.yahoo.com/. The founders Jerry Yang and David Filo compiled a list of interesting pages and after some time these pages were categorised, firstly by categories and then later sub categories. In this way the directory developed organically.

[4] The Google directory can be found at http://www.google.com/dirhp.

new and existing databases, so that data can be shared and reused across a variety of applications. For this architecture to function, however, data must be expressed in a highly structured format and inference rules must be created to enable automated reasoning across a multiplicity of web resources.

1.1.1. Ontologies

Ontologies have been proposed as a foundation for the semantic web. Indeed ontologies, and the field of knowledge representation, are areas that have been studied by Artificial Intelligence (AI) researchers long before the advent of the World Wide Web. The expert system, for example, which emerged from research labs during the sixties and seventies, used a variety of ways to represent real-world knowledge with the aim of conducting very specific domain tasks. The process involved externalising an expert's, or several experts', knowledge and then encoding that knowledge in a machine processable format. The system, when queried, traversed the knowledge base, and formulated a conclusion. If, however, the question is outside of the system scope, it returns a negative answer. Tim Berners-Lee's semantic web proposal operates on similar principles but without the luxury of a limited or closed domain. This is because, unlike the expert system, the web does not represent the knowledge of a single expert, or indeed group of experts, but rather the opinions, suggestions and ideas, whether correct or otherwise, of a billion disparate users.

Nevertheless, in 2004 the World Wide Web Consortium (W3C) recommended both RDF (Resource Definition Framework) (Manola and Miller 2004) and OWL (Ontology Web Languages) (McGuinness and Harmelen 2004) as formal (ontological) languages for semantically representing web resources (W3C 2004). In the same year, Tim Berners-Lee suggested that developers now have the languages to

1. Introduction

This thesis investigates an area of much current research interest: namely the structuring, organising and representation of community-based knowledge. Research in this area is not only concerned with technology; social factors, such as how communities function and thrive, introduce a more multidisciplinary study, while suggesting several outstanding research questions. Notable are the emergence of ontologies, as a way to represent implicit knowledge, and the rise of the social web, as a collective approach to organising user-generated content. It is within this context that this thesis will focus on knowledge representation when approached from a collective or social perspective.

This chapter will firstly introduce some of the background to this thesis. This discussion is rooted in the use of ontologies to structure community-based knowledge, however, narrative is also considered as it is, arguably, one of the most basic ways in which people structure, relate and transfer knowledge. The chapter will go on to introduce the research questions, which are gathered in chapter 4, and presents the statement of contributions, which emerged from this study. The chapter will then introduce the CIPHER project that served as a platform for this research, and finally outline the structure of this thesis.

1.1. Background

In a 2001 Scientific America article, Tim Berners-lee presented the vision of a semantic web, where software agents could digest information and carry out 'sophisticated tasks' for the web user (Berners-Lee, Hendler et al. 2001). The essence of the semantic web, it was suggested, is to expose or 'webise' (Berners-Lee 2001)

List of Figures

Table of Contents

Acknowledgements

I would like to thank Dr. Charlie Cullen for his help and guidance throughout the write-up of this thesis. I would also like to thank Sine, for her endurance, and my family, for their continued support. Finally, I'd like to thank my colleagues and supervisors Eoin Kilfeather and Dr. Ciaran Mc Donnell.

Abstract

The development of ontologies has become an area of considerable research interest over the past number of years. Domain ontologies are often developed to represent a shared understanding that in turn indicates cooperative effort by a user community. However, the structure and form that an ontology takes is predicated both on the approach of the developer and the cooperation of the user community. A shift has taken place in recent years from the use of highly specialised and expressive ontologies to simpler knowledge models, progressively developed by community contribution. It is within this context that this thesis investigates the use of ontologies as a means to representing collective knowledge. It investigates the impact of the community on the approach to and outcome of knowledge representation and compares the use of simple terminological ontologies with highly structured, expressive ontologies in community-based narrative environments.

Declaration

I certify that this thesis which I now submit for examination for the award of ____________________, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for postgraduate study by research of the Dublin Institute of Technology and has not been submitted in whole or in part for an award in any other Institute or University.

The work reported on in this thesis conforms to the principles and requirements of the Institute's guidelines for ethics in research.

The Institute has permission to keep, to lend or to copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature ________________________________ Date ______________

Candidate

A study on the use of ontologies to represent collective knowledge

M.Phil. Thesis

September 2008

School of Media

Dublin Institute of Technology

John McAuley

Research Supervisors: **Eoin Kilfeather, M. Phil.**

Ciaran McDonnell, Ph.D