CRISs, Thesauri and the Semantic Web

Brian Matthews, Alistair Miles, and Michael Wilson
CCLRC, Rutherford Appleton Laboratory

Abstract

For the CRISs to be effective the notion of controlled vocabulary shared by a community is central. Structuring of controlled vocabularies, including taxonomies, thesauri and ontologies has been studied for many years. In this paper we shall investigate the use of one such method of defining controlled vocabularies; thesauri. We compare approaches which have been proposed for expressing such controlled vocabularies in the Semantic Web, and propose a common forms to allow a migration path from existing thesauri to the Semantic Web.

1 Introduction

It has been recognised that a vital component of the infrastructure of CRISs is the use of controlled vocabulary, a language of terms with agreed meanings which can be shared in a particular community. There has been work for many years on structuring controlled vocabularies, including taxonomies, thesauri and ontologies, and they are widely used in CRISs to control resource cataloguing, querying and filtering (Middleton 2000). The World-Wide Web Consortium (W3C) is developing standards for the representation of ontologies to constrain the vocabularies of resource descriptions based on RDF (RDF Schema (Brickley & Guha 2000), OWL (Patel-Schneider et. al. 2003). Such ontologies will allow distributed authoritative definitions of vocabularies to be provided. Such ontology representations are planned to fulfil the role traditionally undertaken by thesauri in CRISs.

This offers an opportunity to use existing work on thesauri to leverage the uptake of the Semantic web, and aid CRISs to be integrated into that Semantic Web. By delivering established vocabularies to the wider community, the use of the semantic markup for web resources can be quickly enabled without additional effort in defining the terminology used. Further, thesauri can then form the basis for developing richer ontological structures. Therefore a migration path is required from current thesauri to the semantic web or support for their co-existence if those ontologies are to be adopted and assimilated into existing information retrieval infrastructure. In this paper we shall compare approaches which have been advocated for expressing thesauri in the Semantic Web, and propose common forms to allow a migration path from existing thesauri to the Semantic Web.

1.1 Knowledge Organisation Systems and Thesauri

Authoritative lists of categorisation terms or controlled vocabularies, generically known as *Knowledge Organisation Systems (KOS)*, have been used in libraries for centuries to catalogue print media. Using terms from a limited controlled vocabulary in searches increases the precision, and when the term is both locatable in the controlled vocabulary and actually used to index documents it will improve the recall (Lancaster 1987). Since the 1970's those word lists have been structured as thesauri to improve the location and selection of terms within and across authorities (Mandel 1987). A thesaurus is a compilation of words and phrases showing synonymous, hierarchical, and other relationships and dependencies, the function of which is to provide a standardised vocabulary for information storage and retrieval systems (Aitchison et. al. 1997).

The structure of thesauri is controlled by international standards that are among the most influential ever developed in the library and information field. The three main standards define the relations to be used between terms in monolingual thesauri (ISO 2788:1986 1986), the additional relations for multilingual thesauri (ISO 5964:1985 1985), and methods for examining documents, determining their subjects, and selecting index terms (ISO 5963:1985 1985). ISO 2788 contains separate sections covering indexing terms, compound terms, basic relationships in a thesaurus, display of terms and their relationships, and management aspects of thesaurus construction, and its general principles are considered language- and culture-independent. As a result, ISO 5964:1985, refers to ISO 2788 and uses it as a point of departure for dealing with the specific requirements that emerge when a single thesaurus attempts to express "conceptual equivalencies" among terms selected from more than one natural language (Austin 1986).

The ISO standards for thesauri (ISO 2788 and ISO 5964:1985) are developed and maintained by the ISO, Technical Committee 46 whose remit is Information and Documentation. ISO 5964:1985 is currently undergoing review by ISO TC46/SC 9, and it is expected that among changes to it will be the inclusion of a standard interchange format for thesauri. To facilitate the growth of the Semantic Web, it would be sensible to try to ensure that such an interchange format is as compatible with Semantic Web ontology representations as possible.

1.2 Thesaurus classes and relations

In order to develop a migration path from current thesauri to Semantic Web representations, it is necessary to understand the structure of thesauri. When searching for information, query terms entered retrieve answers. If the query term is not used to index items then the user needs to know the *preferred* term and to use that instead. If the user receives too few answers they want to broaden the search to recall more items, whereas if the search produces too many answers, they want to narrow the search to produce fewer answers. The hierarchical links in a thesaurus

map onto this desired functionality of *broader* (BT) and *narrower* (NT) search term. Table 1 summarises the standard thesaurus relationships. For any hierarchy there is one and only one Top Term, which can be regarded as a second class of object. The third, fourth and fifth classes are scope notes, dates and histories. Scope notes can be sub-typed into different classes in individual thesauri, but the standard does not do so itself.

In monolingual thesauri, the relation TT exists between any term and its top term in a hierarchy, while BT, NT, RT, UF can exist between two terms. The SN relation exists between terms and scope notes. BT and NT are reciprocal relations on the edges in the hierarchy, BT pointing towards the top term of the hierarchy and NT to the terminal nodes. NT has a relationship akin to child while BT is akin to parent.

To contrast with these *hierarchical* relations, there are *associative* relations: UF, USE and RT. The UF terms is a synonym term with its reciprocal USE, which are used to show that one and only one of a set of terms with equivalent meaning is preferred by the categorisation system and is used for indexing. The referred-from terms include synonyms in direct and inverted word order, alternative spellings (including singular and plural forms), alternative endings, changed or cancelled headings, and abbreviations and acronyms. For phrase headings entered in the inverted form, USE references are made from the straight form. For phrase headings entered in the straight form, USE references are made from the inverted form in selective cases. For compound headings and for topical headings subdivided by other topics, USE references are made from the reversed form, thus bringing each significant term to the initial position. Occasionally, USE references are made to broader headings from narrower terms not used as valid headings. USE references are not generally made from equivalents in foreign languages.

All thesauri		**Multilingual thesauri**
Top Term	TT	Exact Equivalent
Broader Term	BT	Inexact Equivalent
Narrower Term	NT	Partial Equivalent
Related Term	RT	One to Many Equivalent
Used For	UF	Language of
Use	USE	
Scope Note	SN	

Table 1: The relations in all thesauri, with their standard two letter abbreviations and those specific to multilingual thesauri

If non-preferred terms are used in queries for searching, then they should be mapped to the preferred term which has been used for actual labelling. Since terms are words, and terms can occur in multiple hierarchies, it is possible for a word to be the preferred term in one hierarchy (possibly using its major sense) while also being a

non-preferred term in another hierarchy (possibly using a minor sense). In this case the simple use of the word in a query is ambiguous as to which sense is intended. This example shows that there is a notion of concept behind different senses of words within thesauri.
The RT relation is used between two terms that hold an associative relation, but which are not related in the broader/narrower relation of the hierarchy, or through the UF synonymy. Such references may be made for the following types of relationships: headings with meanings that overlap to some extent, headings representing a discipline and the object studied, and headings representing persons and their fields of endeavour (examples: Ships RT Boats and boating; Birds RT Ornithology; Medicine RT Physicians)
It is conventional in multilingual thesauri to have a hierarchy of terms for each language labelled with the language, then to establish relations between individual items across those language hierarchies using one of the four relations.
Given these types and relations are included in the ISO compliant mono- and multilingual thesauri, then any useful Thesaurus Interchange format must include them as well.

1.3 Constraints on the Model

Additionally, we wish to constrain the thesaurus model with extra conditions on the consistency of the thesaurus, summarised as follows.

- Every concept has at least one top concept (not necessarily unique). Thus, without ambiguity, we can treat TopConcept as a partial relation.
- Narrower and Broader concepts are “inverse”. That is, a concept is a narrower concept of its broader concepts and a broader concept of all its narrower concepts.
- Top concepts have no broader concepts.
- Preferred terms are unique to concepts, thus the pair preferred term – language code forms a unique key for a concept.
- Only top concepts have “hierarchy” scope notes. Note that they could be singleton hierarchies.

These constraints cannot be expressed directly in RDF Schema, and points to a OWL based approach best representing thesauri.

2 A Comparison of Different Approaches

Several different approaches have been proposed by different groups to modelling thesauri using RDF Schema or Semantic Web ontology languages such as DAML+OIL and OWL. In this section we describe some of these approaches, giving a categorisation of the different approaches.

2.1 A Term-Based Approach

The most straightforward approach in either RDFS or DAML+OIL/OWL is to model terms as a class of resources, following closely a simple formalisation of the ISO standard. A distinction is made between those terms which are the preferred representation of a concept, and those that are not, usually by creating two subclasses of the overarching term class. Broader/Narrower/Related links between terms are modelled as properties. The domain/range of these properties is then restricted to the preferred-term class, so that these properties can only be used to link members of the preferred term class. The preferred/non-preferred (use for/use) links between non-preferred terms and their preferred alternative are also modelled as properties. These classes and properties are summarised in table 2.
This approach is taken by the Gateway to Educational Materials (GEM), and a fundamentally similar approach is taken by the Dynamics Research Corporation as part of the DAML programme. This latter approach used DAML+OIL and therefore can express more of the constraints on the thesaurus model.
The main strength of this approach is its simplicity; the GEM thesaurus format is particularly straightforward, and it follows the thesaurus standard very closely, accurately models the structure of a monolingual thesaurus. However, for more ambiguous word structures, it is not clear how that it would be satisfactorily extended, especially to the multilingual case as the simple use of preferred term to in the hierarchy does not map well to multiple meanings in different languages. Further, this model does not cope well with conceptual drift where terms change in meaning. This makes it hard to maintain and extend. However, the simplicity of the design does make this approach attractive, especially when using DAML+OIL ontology constructors.

Classes	SubClassOf	Properties	Range
Term	Resource	value	{literal}
Preferred-Term	Term		
		BT	[Preferred-Term]
		NT	[Preferred-Term]
		RT	[Preferred-Term]
		UF	[Entry-Term]
Entry-Term	Term	USE	[Preferred-Term]

Table 2: summary of the RDF classes and properties in the Term-Based approach.

2.2 Concept-Based Approach

In the modelling of thesauri it is tacitly assumed that a group of terms that is the preferred term and its entry terms is being used to describe some *abstract concept*; the text of the ISO standard is rather ambiguous about this point, which is a fine one

for text-based or locally stored thesauri for largely human usage. Strictly speaking, the broader/narrower/related links (the statements about generality) are not being made between terms, but between the concepts they stand for. However, in the term-based approach, each preferred term is taken as a proxy for the concept it stands for, and the broader/narrower/related links are made between these.

In a concept based approach, the concepts are modelled explicitly as a distinct class of resources. Broader/Narrower/Related links are modelled as properties, but the domain/range are Concepts. Each concept is then linked to the terms that can be used to represent it, one which is preferred and any number which are not. The RDF Schema properties and classes for the concept-based approach is given in table 3. This approach has been realised in RDF by (Cross, et. al. 2000), and taken further into the multilingual case by Matthews, Miller and Wilson (Matthews et. al. 2001a; Matthews et. al. 2001b; Miller & Matthews 2001), which was prototyped on the HASSET social science thesaurus (HASSET 1999). This approach is more complex than other approaches with an extra layer of indirection. For example to find the preferred term of a given term, we have to first find the concept of the term via a reverse traversal of the *hasNonPreferredTerm* property, and then traverse the *hasPreferredTerm* property, whilst in a term-based approach this would need a single traversal of the USE property.

Nevertheless, this approach has been advocated by (Doerr & Fundulaki 1998a; 1998b), who argue that this approach solves confusion caused by overloading of terms, where one term can be used for many concepts; in this model this would be reflected directly, with scope notes explaining the qualification, whilst a term-based approach would potentially have the confusion of which meaning is intended. This confusion is exacerbated when the multi-lingual case is considered, as equivalence between terms when there are potentially many alternatives of translation could cause great ambiguity. Further, from an practical view of maintenance, it is more straightforward to maintain systems which evolve over time as the meaning of terms change. For example, this approach allows easy reshuffling of the preferred/ non-preferred terms, without disturbing the generality hierarchy of the concepts.

Classes	**SubClassOf**	**Properties**	**Range**
Concept	Resource		
		classificationCode	{literal}
		hasBroader	[Concept]
		hasNarrower	[Concept]
		IsRelatedTo	[Concept]
		hasPreferredTerm	[Term]
		hasNonPreferredTerm	[Term]
Term	Resource	value	{literal}

Table 3: summary of the classes and properties in the Concept-Based approach.

The concept-based approach captures the intuition that in practical thesaurus construction, the broader-narrower relationship reflects the *extension* of the concept, that is the resources which can be classified under those terms. Thus Doerr and Fundulaki state:
Under this definition, if Descriptor$_A$ has a broader meaning than Descriptor$_B$, then the instance set of the latter is a subset of the former.
Thus in this case, the broader/narrower relationship does after all represent a proper subclass inclusion, but not of the extensions of the terms, but rather the extensions of the concepts.

3 Towards common format for multilingual Thesauri

We make a distinction between concept-based models and term-based models. In a term-based approach, although it may still be tacitly implied that a set of terms represent an abstract concept, the concept is not reified in the model. Terms which are preferred terms become the nodes in the generalisation hierarchy. In concept-based models, it is made explicit that a set of terms is used to represent some abstract concept. One of these terms is the preferred term, the others are non-preferred terms. Broader/narrower/related relations are mode between concepts, not between terms – the concepts are the nodes in the generalisation hierarchy. In a multilingual thesaurus equivalence relations are made between the concepts.
Traditionally thesauri, especially monolingual, have been term based, and most of the schema discussed above follow this tradition. In many cases this would be an acceptable format for thesauri. However, the strength of the un-ambiguity and maintainability of the concept model for thesauri is persuasive, despite the extra complexity the model involves. Thus we propose a update of the RDF Schema developed by Matthews, Miller and Wilson (Matthews et. al. 2001a). This schema is reproduced as an appendix to this paper, whilst the major classes and properties are presented in Figure 1.
Note that the familiar properties of *broaderConcept* and *narrowerConcept* are subproperties of *conceptRelation*, whilst notions of equivalence between concepts, used to define either multilingual thesauri, or thesauri defined from a different domain, are subproperties of the *conceptEquivalence* property.
In this approach, we have taken the view that different versions of the same thesaurus in different languages have their own concept hierarchies, with relations between them. This contrasts with other approaches which take the view that there in one hierarchy with alternate preferred terms for different languages.

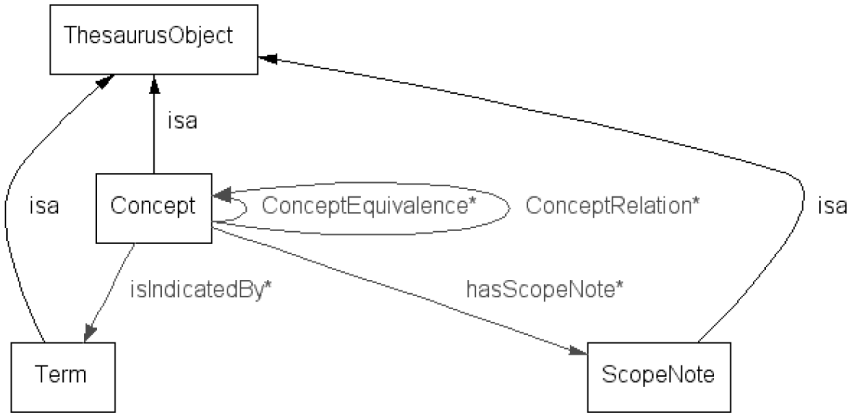

Figure 1: *Simplified multilingual thesaurus model.*

3.1 RDF Schema vs. OWL

The approach taken here in Appendix A is to use RDF Schema. This is to allow a simplicity of expression which can be used immediately be put to use before standardisation of OWL has been completed. However, as has been noted, it would be appropriate to express the thesaurus in OWL, thus allowing constraints, notably the inverse relationship between broaderConcept and narrowerConcept, and the uniqueness of the prefferedTerm, as follows:

```
<owl:Class rdf:ID="Concept">
  <rdfs:subClassOf rdf:resource="#ThesaurusObject"/>
</owl:Class>
<owl:ObjectProperty rdf:ID="hasBroaderConcept">
  <rdfs:subPropertyOf rdf:resource="#ConceptRelation"/>
  <owl:inverseOf rdf:resource="#hasNarrowerConcept"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasNarrowerConcept">
  <rdfs:subPropertyOf rdf:resource="#ConceptRelation"/>
  <owl:inverseOf rdf:resource="#hasBroaderConcept"/>
</owl:ObjectProperty>
<owl:FunctionalProperty rdf:ID="hasPreferredTerm">
  <rdfs:subPropertyOf rdf:resource="#isIndicatedBy"/>
</owl:FunctionalProperty>
```

A further option would be to Concept class itself a subclass of the OWL Class:

```
<owl:Class rdf:ID="Concept">
  <rdfs:subClassOf rdf:resource="#ThesaurusObject"/>
  <rdfs:subClassOf rdf:resource="owl:Class"/>
</owl:Class>
```

This on the would make the move towards converting thesauri into ontologies more explicit. However, their would be a possibility of confusion here. The instances of such concept classes are not terms, but the resources which are classified under those concepts.

3.2 Adding further relationships

In using the concept based approach, we have lost the traditional properties associated with thesauri, namely the relations *BT, NT, UF, USE* and other defined in the ISO standard, which are properties between terms rather than concepts. However, we can reintroduce them into the same OWL thesaurus model, thus allowing:

```
<owl:ObjectProperty rdf:ID="BT">
  <rdfs:domain rdf:resource="#Term"/>
  <rdfs:range rdf:resource="#Term"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="NT">
  <rdfs:domain rdf:resource="#Term"/>
  <rdfs:range rdf:resource="#Term"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="UF">
  <rdfs:domain rdf:resource="#Term"/>
  <rdfs:range rdf:resource="#Term"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="USE">
  <rdfs:domain rdf:resource="#Term"/>
  <rdfs:range rdf:resource="#Term"/>
</owl:ObjectProperty>
```

However, these are not independent properties in their own right, but rather *derived properties*, which should satisfy the following relationships:

$$BT = isIndicatedBy^{-1} \; ; \; hasBroaderConcept \; ; \; isIndicatedBy$$
$$NT = isIndicatedBy^{-1} \; ; \; hasNarrowerConcept \; ; \; isIndicatedBy$$
$$UF = hasPreferredTerm^{-1} \; ; \; hasNonPreferredTerm$$
$$USE = hasNonPreferredTerm^{-1} \; ; \; hasPreferredTerm$$

Where **;** represents relational composition; *(p ; q)(a,c) if and only if there exists some b such that: p(a,b) and q(b,c)*. Note also we do not need to assert that NT and BT are inverses; this is a derived property of inverse nature of hasBroaderConcept and hasNarrowerConcept. However, OWL at present does not have a sufficiently expressive language to express the above equivalences and to derive such consequences.

Other thesaurus format allow other relationships. For example, the DAML+OIL ontology from DRC has a ACK /AF relations between terms for abbreviations and acronyms. The framework we describe is extensible so such properties would be straightforward to add.

4 Conclusions and Future Work

In this paper we have considered the problem of using existing thesauri to support the development of CRISs within the Semantic Web and considered various approaches proposed to effecting this migration, detailing their strengths and weaknesses. From this we proposed an RDF Schema to support concept based thesauri, which is a more appropriate method of modelling thesauri, especially multilingual ones. OWL ontologies and their extensions offer ways of expressing the properties of thesauri in richer fashion; we indicate how this may be achieved.
We have begun to produce tool support for the thesaurus format as part of a set of tools and applications, including a thesaurus server based on web service, and a generic thesaurus based query refinement tool, as well as simple applications built using the RDF based thesaurus.
The next step is to consider the migration from thesauri to an ontology, with a larger range of constraints. This has been considered by some authors, for example (Wielinga 2001). However, in general this is a difficult problem due to freedom in which thesaurus designers have interpreted the standard thesaurus relationships. Ultimately, this may always require human intervention, but ways of assisting the process should be considered, as should ways of mapping between thesauri. However, we believe that in a large number of applications, simple cataloguing and searching of web resources for example, the simple thesaurus structure is likely to prove sufficient.

5 Acknowledgements

We would like to thank colleagues, especially Dan Brickley and Ken Miller. This work supported by the European project Semantic Web Advanced Development in Europe (SWAD-Europe). Further information on the Thesaurus workpackage within SWAD can be found at http://www.w3c.rl.ac.uk/SWAD/thesaurus.html.

6 References

Aitchison, J., Gilchrist, A. Bawden, D. (1997) *Thesaurus construction and use: a practical manual* (3rd Edition) Aslib: London
Amann B. & Fundulaki. I. (1999). *Integrating Ontologies and Thesauri to Build RDF Schemas*. In ECDL-99: Research and Advanced Technologies for Digital Libraries, Lecture Notes in Computer Science, pages 234--253, Paris, France. Springer-Verlag.
Austin, D. (1986), *Vocabulary Control and Information Technology*. Aslib Proceedings 38: 1-15.

Dan Brickley and R V Guha. (2000). *Resource Description Framework (RDF) Schema* Specification 1.0., Candidate W3C Recommendation http://www.w3.org/TR/2000/CR-rdf-schema-2 0000327.
Cross, P., Brickley, D. & Koch T (2000). *Conceptual relationships for encoding thesauri, classification systems and organised metadata collections and a proposal for encoding a core set of thesaurus relationships using an RDF Schema*. http://www.desire.org/results/discovery/rdfthesschema.html
Doerr, M. & Fundulaki, I. (1998a). *SIS-TMS, A Thesaurus Management System for Distributed Digital Collections*, In Proc. of the 2nd European Conf. on Digital Libraries, Heraklion.
Doerr, M. & Fundulaki, I.(1998b) *A proposal on extended inter-thesaurus links semantics*. Technical Report TR-215, Institute of Computer Science-FORTH..
The Gateway to Educational Materials Thesaurus (GEM) http://www.fao.org/agrovoc/
Hall, M. (2001) *CALL Thesaurus Ontology in DAML.* Dynamics Research Corporation. http://orlando.drc.com/daml/ontology/CALL-Thesaurus/G3/CALL-Thesaurus-ont-g3r1.daml,
HASSET (1999). *Humanities and Social Science Electronic Thesaurus.* http://biron.essex.ac.uk/searching/zhasset.html
ISO 5963:1985 (1985) *Documentation -- Methods for examining documents, determining their subjects, and selecting indexing terms*
ISO 5964:1985 (1985) *Documentation--Guidelines for the establishment and development of multilingual thesauri.*
ISO 2788:1986 (1986), *Documentation--Guidelines for the establishment and development of monolingual thesauri 2nd ed.*
ISO 639:1988 (1988) *Code for the representation of names of languages.*
Lancaster, W. F. (1987) *Vocabulary Control for Information Retrieval,* 2nd ed. Washington, DC: Information Resources Press.
Mandel, C. A. (1987) *Multiple Thesauri in Online Library Bibliographic Systems*. Washington, DC: Library of Congress.
Matthews, B.M., Miller, K., Wilson, M.D.,(2001a) *A proposed RDF Schema Thesaurus from the Limber project* http://www.limber.rl.ac.uk/External/ thesaurus-iso.rdf and prototyped using the ELSST social science thesaurus. http://www.limber.rl.ac.uk/External /ELSST_demo_RDF.xml
Matthews, B.M., Miller, K., Ramfos, A., Ryssevik, J., Wilson, M.D., (2001b) *Internationalising data access through LIMBER,* in D.L.Day and L.M.Dunckley (eds) Designing for Global Markets 3: Proceedings of iwips2001, pgs 129-142, Open University: Milton Keynes.
Miller, K., Matthews, B.M. (2001) *Having the right connections: the LIMBER project* , In the Journal of Digital Information 1(8) (http://jodi.ecs.soton.ac.uk/)
Middleton, M. (2000). *Controlled Vocabulary Resource Guide*. http://www.fit.qut.edu.au/InfoSys/middle/cont_voc.html

Peter F. Patel-Schneider, P.F., Horrocks, I., Hayes, P., van Harmelen, F., eds. (2003). *Web Ontology Language (OWL) Abstract Syntax and Semantics* W3C Last Call Working Draft 31 March 2003. http://www.w3.org/TR/owl-semantics/
B.J. Wielinga, A Th Schreiber, J Wielemaker, JAC Sandberg. (2001*) From Thesaurus to Ontology,* K-CAP '01, ACM

7 Contact Information

Brian Matthews, Alistair Miles, Michael Wilson
CCLRC, Rutherford Appleton Laboratory
Didcot
OXON OX11 0QX UK

e-mail:b.m.matthews@rl.ac.uk
e-mail:a.j.miles@rl.ac.uk
e-mail:m.d.wilson@rl.ac.uk