The semantic processing of continuous quantities for discrete terms in ontologies

Shenghui Wang[1] David Rydeheard[1] Jeff Z. Pan[2]
[1]School of Computer Science,
University of Manchester, UK
{wangs,david}@cs.man.ac.uk
[2]Department of Computing Science,
University of Aberdeen, UK
jpan@csd.abdn.ac.uk

Abstract

We consider continuous quantities that are used to describe the physical world, such as colour, shape, sound, texture, and spatial and temporal arrangements. Natural languages are not adept at describing these quantities, nor are they easily incorporated into ontologies in the form of discrete terms. In this paper, we analyse the way that natural languages handle continuous quantities, propose a general semantics based on metric spaces, and describe how to treat semantic values computationally, so that we may automate the processing of texts which describe continuous quantities allowing, for example, query evaluation and the integration of multiple texts. This provides a basis for incorporating these quantities into ontologies and combining their semantics with automated reasoning tools. We begin a series of experiments to evaluate the semantics, the general framework, and the computational system we have developed.

Keywords: Practical semantics, metric semantic model, continuous quantities, similarity, information integration and query evaluation

1 Introduction

There is a fundamental mismatch between the resources of natural languages and one of their important functions, namely, describing the physical world. Natural languages are essentially finite in nature – having a finite (if changing) vocabulary and a finite number of formation rules for finite sentences. The physical world, on the other hand, appears to be continuous. In practical terms, this means that, when describing continuous quantities, we are limited to identifying very few of the values, and, in order to capture the phenomenon of values being arbitrarily close to each other, and eventually indistinguishable, we build in imprecision or under-specification into descriptions of values. The continuous quantities we have in mind include colour, shape, sound, texture, spatial and temporal arrangements, amongst many others. Description of individual values is not the only problem. More challenging still for the resources of language is the description of ranges of variation for these quantities.

In the area of knowledge representation, continuous quantities pose similar difficulties. Concepts associated with words and phrases are represented as discrete terms in an ontology, yet words and phrases for continuous quantities refer to regions of a multidimensional continuum, and these, therefore, require a treatment in which a normal ontological description is combined with appropriate properties of the continuum.

The processing of words and phrases for continuous quantities raises important issues in the treatment of semantics – issues that we address in this paper. The problem becomes particularly focused when we consider its computational aspects. In the automated analysis of text, a full processing of queries to match them to text so as to retrieve information is based, in general, on semantics, rather than lexical or syntactic information. Similarly, comparing texts to assess the amount of agreement, overlap, consistency and inconsistency, and the related process of integrating multiple texts into a single account without duplication but detecting inconsistency, requires, for its full implementation, a computational semantics [SvH04]. To incorporate continuous quantities into ontologies, or other formal knowledge representations, and into description logics also requires a semantics in an appropriate form.

One area where the language of continuous quantities is particularly important is that of the descriptive sciences, such as botany (our primary example), anatomy, zoology etc. In these areas there are large corpora of written knowledge. In botany, in particular, there is a vast collection of *parallel* (and often independent) descriptions with the same species described by many different authors. Expressions for continuous quantities form a major component of these descriptions, for example, expressions for flower colour, leaf shape, arrangement of parts, vein patterns, etc. and also a wealth of temporal and spatial expressions. These written descriptions have a degree of precision and detail not normally encountered in texts.

The problem of automatically comparing and integrating information is especially difficult for descriptions of natural subjects because there is variation not only because different authors use different words and phrases, but also because nature itself is highly variable in such a way that no simple phrases can capture the full range of variation and different authors may approximate in different ways. As a simple example, the flowers of *Origanum vulgare* (Marjoram) are described by several authors as:

- 'violet-purple'
- 'reddish-purple, rarely white'
- 'white or purplish-red'
- 'purple-red to pale pink'

Whilst we can appreciate, at a glance, to some extent the degree of similarity and disagreement amongst these authors, an automated process to do this makes heavy demands on the formulation of a usable semantics for each continuous quantity.

In this paper, we provide methods for processing phrases describing continuous quantities, by introducing (1) a general semantics of such phrases together with methods of measuring the similarity between denotations, (2) appropriate models for several such quantities, (3) a formal ontology language which is expressive enough to represent these quantities, and (4) an improved, semantic, treatment of query matching and a strategy for integrating information from different sources. Finally, we begin a series of experiments to evaluate the framework, both the semantics and the computational techniques.

Whilst the emphasis of this work is very practical, we note several features of general linguistic interest: The first is that we have undertaken an extensive automated analysis of the types of phrases which are used to describe several continuous quantities. What this reveals is that there are actually only a few general ways of forming phrases for continuous quantities in English and, in the main, these are uniform across all such quantities. Moreover, we can describe a uniform semantics for these. That is, the way we interpret such phrases is independent of the quantity under consideration. Of course, to give a semantics for a particular continuous quantity, we need to choose a suitable model, but then the interpretation of phrases is determined by specialising a general account to this particular model.

2 Approaches to semantic analysis

Many of the issues arising in the semantics of natural languages (see, for example, [Lyo95]) are not relevant here, such as those dealing with full sentential meaning in terms of verb and noun phrases, tense, voice, ellipsis etc. We are dealing with a restricted class of noun and adjective phrases, and their conceptual content is not usually at issue. However, some of the issues, such as what exactly is a denotation, how do we handle the imprecision, ambiguity and under-specification of natural languages, what is the meaning of individual descriptive words, and how do we validate definitions of meaning, are relevant here.

For some continuous quantities, such as colour, there has been considerable work done, both in linguistic aspects of colour, for example in the concept of 'language universals', and also in the very practical area of interpreting colour names in terms of colour models, see [BBK82, Moj02, Tom85] amongst many papers in this area. Some of the latter are of direct use in the semantics that we propose, providing a means of interpreting basic terms from which phrases are formed.

The semantics we describe is presented in terms of metrics. Similar metric structures have been employed in areas of cognitive science, see, for example, [GÖ0]. Spatial or geometrical structures are exploited in concept formation and learning, and also in studies in cognitive linguistics [Lak87].

A key aspect of the proposed semantics is the calculation of semantic distances and similarity [SJ99]. Various general methods have been proposed, including the ratio of common/distinct features in *feature models* [Tve77], the vector distances in multi-dimensional *spacial models* [OST57, GÖ0, LFL98], the path-length in *network models* [RMBB89, WP94], etc. In natural language research, corpus-based methods are used to measure similarities between concepts by comparing their information content [Res95].

Despite the wealth of work on natural language semantics, many of the very specific problems associated with interpreting and processing phrases for continuous quantities are not resolvable with these general techniques, and whilst some of the work on interpreting particular classes of words for continuous quantities (such as colour) are relevant here, we are not aware of any successful attempts to extend these to general phrases for continuous quantities.

3 A metric semantics

We now turn to a general semantics of expressions for continuous quantities, interpreting expressions in metric spaces. A metric space is a set X together with a 'distance measure' (called a *metric*) $\mu : X \times X \rightarrow \mathbf{R}^+$ (where $\mathbf{R}^+$ is the set of non-negative real numbers) with μ satisfying, for all $x, y, z \in X$,

1. $\mu(x, y) = 0 \iff x = y$,
2. $\mu(x, y) = \mu(y, x)$,
3. $\mu(x, z) \leq \mu(x, y) + \mu(y, z)$.

A metric space provides a notion of closeness or proximity of points, and, in terms of this proximity, we may define regions of the space.

Important aspects of metric spaces which we exploit in the semantics are (1) paths in spaces and, in particular, the notion of a shortest path between two points, often called a *geodesic*, (2) definitions of continuity of maps between spaces, and (3) various products of spaces. For a summary of those areas of metric space theory motivated by this development of semantics, see [RW07]. For example, well known metrics, over subsets of $\mathbf{R}^n$, include (1) the so-called *city-block* metric defined by:

$$\mu((x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n)) = \sum_{i=1}^{i=n} |x_i - y_i|,$$

and (2) the Euclidean metric:

$$\mu((x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n)) = \sqrt{\sum_{i=1}^{i=n} (x_i - y_i)^2}.$$

As we will show later in this section, we mainly consider regions in metric spaces, so we will propose some region-based metrics in Section 3.3.

Metric spaces may not be the only setting for a semantics of continuous quantities in natural languages, but it is the purpose of this paper to show that they provide workable models of sufficient generality and useful applicability.

3.1 Denotations

Points in a metric space are the location of exact values. For continuous quantities, as they occur in nature or as described in natural languages, collections of points rather than single points are the appropriate denotation for phrases, allowing points close to and indistinguishable from each other to be collected together in the denotation. This spread of meaning, arising from the imprecision of language, leads us to interpretations of words and phrases as *regions* in a metric space. Another reason for considering regions is that some expressions denote ranges of values, either around a point, or between two or more points, and the natural interpretation of these is as regions of the space.

Determining the location, extent and shape of regions which correspond to phrases is, however, often difficult and raises a number of general issues about natural language semantics. The key problem is the under-specification that is present in our use of language. This vagueness which occurs in most description, even for texts in the descriptive sciences, means that regions are rarely determined precisely. Definitions of terms (if they exist at all) usually describe values – often called *prototypical* values – rather than regions in a modelling space. The move from such points to definitions of appropriate regions around them is often not obvious, and yet it is crucial to the question of the similarity and overlap of meanings of phrases. Several approaches to this have been suggested, including using a fixed, or varying, granularity for the denotations (possibly using 'Voronoi diagrams'), or alternatively, using fuzzy sets and 'graded membership'.

Computing with regions as denotations rather than points makes for additional complication in the definition of semantics. For example, how do we determine the similarity between semantic values when these values are regions? Various notions of similarity between regions in metric spaces have been proposed [VH01]. Four approaches can be identified: (1) choosing a co-ordinate frame and comparing regions in each co-ordinate separately, (2) measures of distance between regions defined in arbitrary metric spaces, e.g. Hausdorff distances, (3) optimal positioning – determining the translation and rotation (and possibly expansion) needed to optimally place one region onto another, and (4) 'morphing' techniques – determining a mapping to make one region exactly match another. In this paper, we concentrate on method (1) and in Section 3.3 show how to calculate a distance between regions in terms of co-ordinates determined by the dimensionality of the continuous quantity in question. We then use this distance on regions to determine the outcome of queries and for an information integration technique.

3.2 Descriptive phrases for continuous quantities

In order to determine the phrases in use in English for continuous quantities, we have chosen two quantities, colour and a simple form of shape (leaf shape), and processed botanical texts to extract the syntactic forms of these phrases. The texts used are so-called floras,[1] written with a mixture of technical terms, numerical expressions and standard English. The only technical terms we use in this paper are leaf-shape terms. These may be unfamiliar to the reader so we illustrate some of them in Figure 1. The automated analysis was carried out using 227 colour descriptions and 362 leaf-shape descriptions from five floras.[2] In Table 1, we list the forms of expressions for continuous quantities which occur in these texts.

We now explain how these phrase formations are to be interpreted in an arbitrary metric space. Later, we will choose suitable modelling spaces for colour and for leaf shape.

Basic terms

We begin with the denotations of single-word basic terms. Where do these come from? The shape, location and extent of basic regions clearly depends on quantity being modelled and the modelling space as well as the precision of the descriptive term.

[1] Floras contains descriptions of groups of plants often in a particular geographic region.

[2] They are *Flora of the British Isles* (Clapham, Tutin and Warburg, 1987), *New Flora of the British Isles* (Stace, 1997), *Flora Europaea* (Tutin, Heywood, Burges, Valentine and Moore(eds), 1993), *The Wild Flower Key* (Rose, 1981) and *Gray's Manual of Botany* (Fernald, 1950).

Syntactic form	Examples	
	Colour	Leaf shape
Single term	‘purple’	‘ovate’
Modified terms	‘pale pink’	‘broadly elliptic’
Hyphenated expressions	‘greenish-yellow’ ‘blue-purple’	‘linear-lanceolate’
Ranges built using ‘to’	‘blue to violet’	‘oblong to elliptic’
Multiple expressions with connectives (‘and’, ‘or’), and/or punctuation	‘lavender, white-pink’ ‘white and green’	‘linear or narrowly elliptic’ ‘ovate and cordate’
Mixed expressions	‘violet-blue, rarely pink or white’	‘ovate, often deltoid’

Table 1: Syntactic forms of expressions for continuous quantities

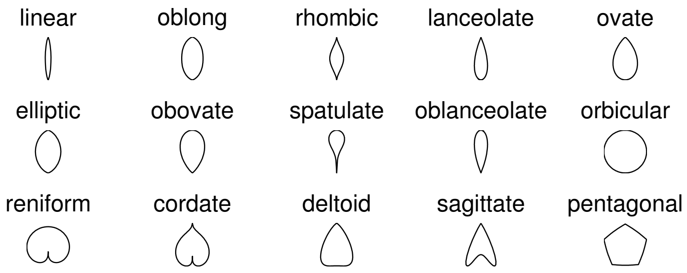


Figure 1: Some example leaf shapes (with the base at the bottom of the image)

For some quantities, there are some systematic descriptions of basic terms via *naming schemes*, such as, [BBK82, Tom85] for colours. However, in general, there are no established definitions for basic terms. Sometimes *glossaries* exist which describe the meaning of basic terms usually fairly imprecisely. For example, for leaf shapes, some texts use typical pictures, others, such as [Sta97] and [MM98], attempt to delimit the meaning of terms (differently!). In such cases, we are able to convert the descriptions, using domain knowledge, into regions in a metric space, and then evaluate the constructed semantics using methods we describe later.

Ranges and intermediates

Consider phrases such as, for colour, ‘pink to purple’ or, for leaf shape, ‘ovate to elliptic’. These denote ranges of values. A natural interpretation of these is in terms of paths in a metric space. Which paths? The relevant paths for these expressions are the *shortest paths*, also known as *geodesic* paths, in the space. That these provide correct interpretations may appear surprising and something of an arbitrary choice. However, in the construction of models for various quantities, metrics are determined by the semantic role of the models, reflecting the way that we perceive and communicate differences between values. With such a choice of model and metric, geodesic paths do indeed cover the possible interpretations of expressions for ranges.

We need to extend the notion of geodesic paths between points to geodesic paths between regions, resulting itself in a region of the space. The appropriate region is that determined by the collection of all geodesic paths between all pairs of points in the

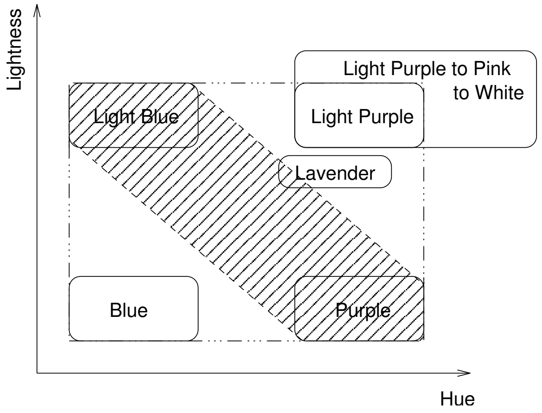


Figure 2: Possible interpretations of 'light blue to purple'

source and target regions. The shape of this region and its properties depend upon the metric. In Figure 2, we show several possible interpretations of 'light blue *to* purple':

- light blue to purple directly (the shaded area),
- light blue to blue then to purple,
- light blue to light purple then to purple,
- the whole rectangle.

The first corresponds to the Euclidean metric in this space, the second and third are geodesics in the city-block metric, which cover the whole rectangle.

A related form of phrase in common use is 'violet-blue' or, for leaf shape, 'ovate-lanceolate'. These compounds again denote regions of the space. In this case, the regions are determined by points along geodesic paths, that is, the phrases denote intermediate values in the ranges determined by the end-points. Exactly which points along the geodesic depends upon the model: half-way along is appropriate for some models but not for others. Indeed, some terminological schemes for continuous quantities (including for colour and for leaf shape) may specify the location of the intermediate values.

Modifiers

Consider phrases such as 'pale yellow' or 'deep reddish purple', or, for shape, 'broadly elliptic' or 'narrowly linear'. Each of these uses a *modifier*, such as 'pale' or 'broadly', to change the denotation.

Modifiers are interpreted as *continuous transformations* of a metric space into itself. The continuity here reflects the fact that physical indistinguishability is a coarser relation than limiting proximity in models. Other preservation properties (e.g. isometry) may hold, but are not expected in general. Which continuous transformations correspond to individual modifier terms is, of course, part of the interpretation of basic terms which we discuss above.

Logical operations

Logical operators naturally occur in the description of continuous quantities and the complexity of their variation. Disjunction is common, for example, 'obovate to narrowly elliptic or narrowly obovate', or 'violet-blue, rarely pink or white'.

In our automated extraction of phrases appearing in botanical texts, we find that the word *and* is rarely used as a simple conjunction of values, but instead occurs in contexts such as, 'narrowly spatulate to linear-spatulate and round' combining different quantities, or in attempting to indicate where variation occurs, for example, 'with numerous oblong and elliptic small lenticels' (with *and* often indicating that on individual specimens two or more values coexist). Negation is rare and almost entirely confined to distinguishing between close taxa.

Logical operations of disjunction, conjunction etc. combine regions of space. For disjunction, the standard Boolean interpretation as the union of regions is appropriate for the models we have considered.

Another logical aspect of the semantics is the presence of *subsumption*, that is, some phrases subsume others. For example, flowers which are 'golden' are also 'yellow' (of course!). This is a simple example, but other instances of subsumption are not so obviously detectable. The presence of subsumption is important when we wish to compare, or to integrate, different descriptions of the same species. Standard subset inclusion provides a suitable interpretation of subsumption.

Even with this fairly restricted use of logic, we need to consider how the semantics may be combined with automated reasoning tools, for example, in description logics. We describe later a system we have built, based on this semantics, for query evaluation and for automated information integration in which we do indeed incorporate tools to support deduction.

Mixed forms

Another form of expression for continuous quantities is common: those which mix several types of quantity. These occur especially in trying to describe the complex variation that is often present in nature, where the variation of one quantity depends upon others.

Examples of this are (combining colour and frequency) 'violet-blue, rarely pink or white', or (combining shape and frequency) 'usually broadly ovate, occasionally deltoid', or (combining position with shape) 'lvs oblong-ovate to lanceolate, narrowing up the stem', or (combining position and size) 'lower lvs large, upper smaller'. Each of these combines two quantities (indeed it would be possible to combine colour and shape – e.g. when petals display different shapes with different colours).

To interpret such phrases, we consider pairs of semantic objects as though they were themselves semantic objects. That is, the interpretation takes place in a *product metric space*. Various products are available reflecting the fact that the metrics for the two quantities can be combined in a variety of ways, which allows us to capture the appropriate interaction of the quantities in the semantics.

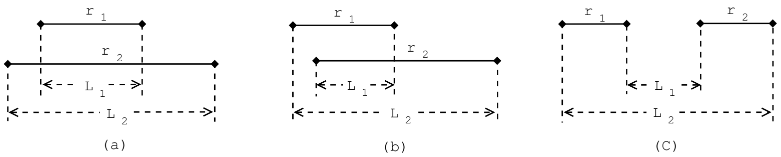


Figure 3: Three relations between two ranges

Ambiguity, imprecision and under-specification

Natural languages expressions, even those appearing in the technical literature of the descriptive sciences, are inherently imprecise in their interpretation, not just in which values are denoted, but also in the extent of the range of semantics values. Although glossaries, definitions and naming schemes may be available for some terms and phrases, the natural language context of these and the imprecision of authors in their use of terms, always mean that there in an inherent under-specification of the intended semantics. One of the fascinating aspects of this mathematical investigation of the semantics and its application in computational contexts is the way that the precision of the mathematics and computation interacts with, and exposes, the precision and imprecision of the use of phrases in the texts.

Another aspect of natural languages is ambiguity. Complex variation of continuous quantities in nature lead to complex expressions, often in need of disambiguation. Even a simple example such as 'pink to purple or blue' is open to two interpretations : either 'pink to (purple or blue)' or '(pink to purple) or blue'. More complex examples often admit several sources of ambiguity, in particular determining exactly where variation occurs: within a single structure, within individuals, or within groups.

3.3 Semantic distance

We now consider the question of measuring the semantic similarity or distance between regions in a metric space. In multi-dimensional spaces (i.e. subsets of $\mathbf{R}^n$), simple regions are determined by intervals in each dimension, that is n-dimensional rectangles. Whilst this is a limited notion of region, it corresponds to the simple semantic idea that a region is determined by the spread of values in suitably chosen dimensions.

How do we determine an appropriate measure of distance between rectangles? Consider the case of $n = 1$, i.e. bounded closed intervals $[a, b]$ in $\mathbf{R}$ with $a < b$. Figure 3 shows the possible relative positions of two such intervals $r_1 : [a_1, b_1]$ and $r_2 : [a_2, b_2]$.

One possibility is to use the *Hausdorff metric*:

$$\mu_H(A', B') = max\{\sup_{x\in A'} \inf_{y\in B'} \mu(x, y), \sup_{y\in B'} \inf_{x\in A'} \mu(x, y)\}$$

For intervals, this is

$$d_H(r_1, r_2) = max\{|a_2 - a_1|, |b_2 - b_1|\}. \tag{1}$$

Another possible measure of distance is similar to the feature contrast model proposed in [Tve77]:

$$d_C(r_1, r_2) = \begin{cases} L_2 - L_1 & \text{if } r_1 \text{ and } r_2 \text{ overlap,} \\ L_2 + L_1 & \text{otherwise.} \end{cases} \tag{2}$$

where L_2 is the length of smallest interval containing r_1 and r_2, and L_1 is defined as follows: when r_1 and r_2 overlap (see (a) and (b)), L_1 is the length of the overlap; otherwise, for (c), L_1 is the length of the gap between them. If two ranges r_1 and r_2 only share one point, we say they *meet* each other and $L_1 = 0$. It is easy to show that (2) is equivalent to the following function $d_C(r_1, r_2) = |a_2 - a_1| + |b_2 - b_1|$, which could be interpreted as the city-block metric on the endpoints of the intervals.

A refinement of this is to relativise the distance by L_2:

$$d_{RC}(r_1, r_2) = \begin{cases} 1 - L_1/L_2 & \text{if } r_1 \text{ and } r_2 \text{ overlap,} \\ 1 + L_1/L_2 & \text{otherwise.} \end{cases} \tag{3}$$

This is bounded in the range $[0, 2)$. Whilst d_{RC} may appear somewhat arbitrary we shall show later that it provides a simple measure which corresponds well to semantic distance, and, moreover, appears to be of some independent mathematical interest.

To extend these measures on intervals to measures on n-rectangles in (subsets of) $\mathbf{R}^n$, we take the (possibly weighted) sum of the measures of the projections onto each axis, with the weightings reflecting the different role or *salience* [GÖ0] of each dimension. Again, there is a semantic justification for this sum in terms of the way each dimension independently contributes to the amount of separation between regions, and we show later that this does provide a useful working measure of the degree to which two denotations are similar.

3.4 Defining a semantics

We now give details of how to define the semantics of descriptions of a particular quantity. The starting point is to choose an appropriate multi-dimensional model, i.e. a subset of $\mathbf{R}^n$ with an appropriate metric. We consider two examples from botanical texts, (1) flower colour descriptions, and (2) leaf shape descriptions.

For colour, an appropriate model is the Hue-Saturation-Lightness (HSL) model. This classifies colours by three aspects which are meant to represent how we perceive, analyse and match colours. The three-dimensional space consists of polar co-ordinates (s, h) with the hue h representing the angle around the ‘colour-wheel’, and the saturation s the distance from the centre of the wheel. Lightness is orthogonal to this and makes the colour-wheel into a double cone with apices being black (lightness is 0) and white (lightness is 100) (see [FDFH90] for a discussion of colour models). The appropriateness of this model lies in its attempt to define axes which appear to correspond to some degree with our perception of colour.

For leaf shape, there is a variety of possible models, from approximation and curve-fitting techniques, through curve-generating formulae, to transformational techniques. We consider here a simple model in which we choose several features of a leaf shape to provide a multi-dimensional description. As an example, we choose four features: the length/width ratio, the position of the broadest part, and the angles at the apex/tip and at the base. Thus the modelling takes place in a 4-dimensional space.

We are now in a position to interpret descriptive phrases in these models. We consider here the city-block metric, where both basic regions and those determined by collections of geodesic paths are n-dimensional rectangles.

1. We begin with a collection of basic terms for the quantity, together with their synonyms and inflections. To each basic term, a *prototypical* point, chosen to

represent its semantics, is allocated. For example, an appropriate position in HSL for ‘purple’ is (83, 50, 25).

2. Regions for simple words or phrases are generated as follows:

   Single terms: A basic term corresponds to a region of the semantic space generated from the prototypical points $P = (p_1, \ldots, p_n)$, where $p_i \in \mathbf{R}$. There are various schemes for determining appropriate regions. The simplest is to give a small range in each dimension, defining a region as $R = (r_1, \ldots, r_n)$, where $r_i = [p_i - \epsilon, p_j + \epsilon]$, with ϵ suitably determined by the scale of the modelling in each dimension and by the appropriate spread of values. For example, ‘purple’ may represented as (78-88, 45–55, 20–30) using $\epsilon = 5$ in each dimension. Whilst there is a degree of arbitrariness about this choice, we shall see that it is well within the imprecision of language and it provides successful mechanisms for semantic processing. Indeed, there is a feedback mechanism allowing us to ‘tune’ these choices (and those below) by comparison with the results of the semantic processing.

   Modifiers: For modifiers, we describe continuous transformations of the space, each determined by the meaning of the modifier. For example, ‘pale’ means fix the hue, but decrease the saturation and increase the lightness, both by fixed amounts. Therefore, ‘pale purple’ transforms ‘purple’ (78-88, 45–55, 20–30) into the region (78–88, 25–35, 40–50).

   Hyphenated expressions: A hyphenated expression ‘X-Y’ determines an intermediate value on the geodesic paths from X to Y. For example, ‘blue-purple’ is generated from the halfway colour between ‘blue’ (66, 100, 50) and ‘purple’ (83, 50, 25), that is, the colour with HSL value of (75, 75, 38) representing the region (70–80, 70–80, 33–43).

   For colours, there is another form of expression: ‘Xish-Y’, which we interpret as a quarterway value on geodesic paths from Y to X, closer to the Y [BBK82].

3. We combine the above simple regions to construct regions of more complex descriptions, as follows:

   Single ‘to’ range: Using the city-block metric, regions determined by ranges defined in terms of ‘to’ are n-dimensional rectangles. For regions connected by one or more ‘to’s, the resulting region is the whole range from the first one to the last. That is, the range which covers two regions $R_1 = (r_{11}, \ldots, r_{1n})$ and $R_2 = (r_{21}, \ldots, r_{2n})$ is $R^{'} = (r^{'}_1, \ldots, r^{'}_n)$, where $r^{'}_i$ is the smallest interval including r_{1i} and r_{2i}.

   Multiple regions: If regions are connected by any of the following symbols: ‘or’, ‘and’, comma (‘,’) or slash (‘/’), we do not merge into single regions, but instead use the collection of individual regions as the denotation. Thus denotations are, in general, collections of n-dimensional rectangles. This corresponds to the semantic idea that these connectors each represent a disjunction. Notice that, as discussed above, ‘and’ is also treated as a disjunction.

Using a simple parser, we can now compute the semantics of a complex description formed from the components above as one or several regions in a metric space.

4 Representing continuous quantities in an ontology

In this section, we describe how to represent formally the semantics described in the last section. We discuss the language features required and show how to represent our example quantities, colour and shape, in an existing ontology system.

4.1 Ontology and its language

As a shared formal understanding of application domains, ontologies are a practical knowledge representation used in many information systems. Formal languages for ontologies have been developed extensively, among which the Web Ontology Language OWL [BvHH$^+$04] is a W3C recommendation for expressing ontologies in the Semantic Web. We use one of its main sub-languages, OWL DL, to build a suitable domain knowledge base. An OWL DL ontology consists of a set of *axioms*, which represent concepts, relationships between concepts and instances. There are associated automated reasoning tools. OWL DL is not quite sufficient to represent the semantics introduced in this paper. In particular, it does not support user-defined datatypes, especially, for our purpose, delimited ranges of data values.

To solve similar problems, Pan and Horrocks [PH05] proposed OWL-Eu, an extension to OWL DL which supports customised datatypes through unary datatype expressions (or simply datatype expressions) based on unary datatype groups.[3]

Let $\mathcal{G}$ be a unary datatype group. The set of *$\mathcal{G}$-datatype expressions*, $\mathbf{Dexp}(\mathcal{G})$, is inductively defined in abstract syntax as follows [PH05]:

1. *atomic expressions*: if u is a datatype symbol, then $u \in \mathbf{Dexp}(\mathcal{G})$;
2. *relativised negated expressions*: if u is a datatype symbol, then $\texttt{not}(u) \in \mathbf{Dexp}(\mathcal{G})$;
3. *enumerated datatypes*: if $l_1, \ldots, l_n$ are literals, then $\texttt{oneOf}(l_1, \ldots, l_n) \in \mathbf{Dexp}(\mathcal{G})$; with arity 1, where $\{\}$ is called the `oneOf` constructor;
4. *conjunctive expressions*: if $\{E_1, \ldots, E_n\} \subseteq \mathbf{Dexp}(\mathcal{G})$, then $\texttt{and}(E_1, \ldots, E_n) \in \mathbf{Dexp}(\mathcal{G})$;
5. *disjunctive expressions*: if $\{E_1, \ldots, E_n\} \subseteq \mathbf{Dexp}(\mathcal{G})$, then $\texttt{or}(E_1, \ldots, E_n) \in \mathbf{Dexp}(\mathcal{G})$.

OWL-Eu ontologies also contain a set of axioms for classes, properties and individuals. FaCT-DG [Pan04], a datatype group extension of the FaCT ontology reasoner, supports reasoning in OWL-Eu ontologies that do not contain nominals. The reader is referred to [PH05] for more details of datatype expressions and unary datatype groups.

4.2 Representing colour and shape

We consider flower colour and leaf shape to illustrate how we represent the descriptions of these quantities in an ontology system, using the OWL-Eu ontology language. The construction rules introduced in the previous section enable us to express the semantics as (collections of) regions in particular metric spaces, which we represent as DL datatype expressions.

Our plant ontology $\mathcal{O}$ contains the knowledge of plant parts and their morphological features. The following primitive classes are concerned in this paper:

[3] A unary datatype group is a set of primitive datatypes and its sub-datatypes built using parameterised datatypes and unary datatype expressions.

Class($\mathsf{Species}$), Class(Flower), Class(Colour), Class(Leaf), Class($\mathsf{LeafShape}$);

important object properties related to them include

ObjectProperty($hasPart$), ObjectProperty($hasColour$),
ObjectProperty($hasShape$);

and important datatype properties include

$hasHue$, $hasSaturation$, $hasLightness$,
$hasLengthWidthRatio$, $hasBroadestPosition$, $hasApexAngle$ and $hasBaseAngle$,

which are all *functional* properties. Each datatype property and its range is also defined, for example,

DatatypeProperty($hasBaseAngle$ Functional range(and($\geqslant 0, \leqslant 180$))).

Typical relations between classes include:

$$\begin{aligned}
&\mathsf{Species} \sqsubseteq \exists hasPart.\mathsf{Leaf} \sqcap \exists hasPart.\mathsf{Flower} \ \ \text{(Each species has leaf and flower parts)}\\
&\mathsf{Leaf} \sqsubseteq \exists hasShape.\mathsf{LeafShape} \ \ \text{(Each leaf has a property: LeafShape)}\\
&\mathsf{Flower} \sqsubseteq \exists hasColour.\mathsf{Colour} \ \ \text{(Each Flower has a property: Colour)}
\end{aligned}$$

Concrete colours and leaf shapes are defined based on the above primitive classes and properties, where datatype expressions are used to express the semantic regions. For example, the colour 'purple' and the shape 'ovate' are defined as the following OWL-Eu classes:

$$\begin{aligned}
\mathsf{Purple} \equiv\ & \mathsf{Colour}\ \sqcap\\
& \exists hasHue.(\geqslant 78\ \sqcap \leqslant 88)\ \sqcap\\
& \exists hasSaturation.(\geqslant 45\ \sqcap \leqslant 55)\ \sqcap\\
& \exists hasLightness.(\geqslant 20\ \sqcap \leqslant 30)
\end{aligned}$$

$$\begin{aligned}
\mathsf{Ovate} \equiv\ & \mathsf{LeafShape}\ \sqcap\\
& \exists hasLengthWidthRatio.(\geqslant 15\ \sqcap \leqslant 18)\ \sqcap\\
& \exists hasBroadestPosition.(\geqslant 39\ \sqcap \leqslant 43)\ \sqcap\\
& \exists hasApexAngle.(\geqslant 41\ \sqcap \leqslant 50)\ \sqcap\\
& \exists hasBaseAngle.(\geqslant 59\ \sqcap \leqslant 73)
\end{aligned}$$

Finally, a species with 'purple' flower and 'ovate' leaf shape will be represented as a OWL class as follows:

$$\begin{aligned}
&\mathsf{SpeciesA} \equiv \exists \mathsf{Species} \sqcap hasPart.\mathsf{LeafA} \sqcap \exists hasPart.\mathsf{FlowerA} \sqcap \ldots\\
&\mathsf{LeafA} \equiv \mathsf{Leaf} \sqcap \exists hasShape.\mathsf{Ovate} \sqcap \ldots\\
&\mathsf{FlowerA} \equiv \mathsf{Flower} \sqcap \exists hasColour.\mathsf{Purple} \sqcap \ldots
\end{aligned}$$

Similarly, species with complex flower colour and leaf shapes are also defined as OWL classes, with flower colour and leaf shape represented as OWL-Eu classes.

Ontological representations of colour and shape descriptions, together with appropriate distance functions for each property, enable us to integrate parallel descriptions and to carry out species identification queries based on flower colour and/or leaf shape.

5 Distance-based integration

The first application is to the integration of multiple parallel descriptions. As above, we exemplify with colour and leaf shape.

	ρ		
	d_{RC}	d_C	d_H
Expert 1	0.57479	0.54843	0.54843
Expert 2	0.65676	0.37897	0.37897
Expert 3	0.53061	0.47866	0.47866
Expert 4	0.26041	0.23770	0.23770
Average judgement	0.63587	0.52528	0.52528

Table 2: Spearman rank correlation coefficient between experts judgements and different distance functions, where the degrees of freedom is 48 and the critical value with the probability of $p = 0.01$ is 0.354

5.1 An integration strategy

The problem is to determine when different descriptions are sufficiently close to be integrated into a single description, and then to perform a semantic integration. If parts of descriptions are not sufficiently similar, then they ought to be kept separate on the understanding that the descriptions actually contain different information which it is not correct to merge into a resultant single description.

The trigger for the integration to take place is based on one of the semantic distances on regions described in Section 3.3 together with a *threshold*. To determine a suitable threshold, a group of parallel descriptions was selected, not identical yet sufficiently similar to be combined. The average distance of these parallel descriptions was used as the threshold. Again this is open to refinement through feedback from the results of integration experiments.

In processing a botanical text, for each species, different continuous quantities are processed separately. A special-purpose reasoner has been devised to carry out the following iterative integration process on the collections of regions determined from parallel descriptions.

Step 1 Calculate the distances between any pair of regions.

Step 2 Select the pair consisting of the two closest regions, and check whether they are similar enough, i.e. whether their distance is less than the threshold. If they are not similar enough then the integration stops; otherwise, the smallest region containing both of them is generated (this is same operation as building 'to' ranges in Section 3.4). This new region replaces the original two as their integrated result.

Step 3 Go back to Step 1 to check the updated collection of regions to see whether there are any further regions requiring integration.

The result of this process is a collection of regions together with the relationship between the original parallel descriptions and the integrated regions. Such relationship, including equality, subsumption, (non-empty) intersection and disjunction, determined by the FaCT-DG DL reasoner.

5.2 Validation tests for distance functions

Before turning to the above integration process, we need to consider what distance measure on regions to employ. We have undertaken an experiment using leaf-shape

Species	Leaf Shape Descriptions	Integration Results			
		R_{f_1}	R_{f_2}	R_{f_3}	R_{f_4}
Salix pentandra (Laurel willow)	ovate or ovate-elliptical to elliptical- or obovate-lanceolate broadly lanceolate to ovate-oblong broadly elliptical broadly lanceolate, ovate-oblong, or elliptic-lanceolate	121–287	27–57	10–35	27–37
Spinacia oleracea (Spinach)	hastate to ovate	122–163	8–39	17–25	37–63
	ovate to triangular-hastate	181–221	45–55	27–33	27–33
	oblong				

Table 3: Examples of integration of parallel leaf shape descriptions, where R_{f_i}, $i = 1 \ldots 4$ are the ranges in four features of leaf shape

expressions as an example. From 21 basic leaf shape terms, we randomly selected 50 triples, each of which consists of a target term T and two candidate terms A and B. Using a distance measure, we can tell whether the distance $d(T, A)$ is smaller than $d(T, B)$, namely, whether A is more similar to T than B is. These triples were presented to four experts – authors of floras. For each triple, they were asked independently to answer which term, A or B, is more similar to term T.

Spearman rank correlation coefficient was calculated to measure the degree to which the expert judgements and the judgement from the calculation are related. As shown in Table 2, the distance function d_{RC} brings our calculations closer related to the expert judgements, for all experts, than the other two functions, the distance function d_{NC} was adopted for the following integration and querying experiments.

5.3 Integration experiment

To evaluate the integration strategy and the semantics on which it is based, we have selected 656 species from five floras together with the online eFloras.[4] Each species has at least two descriptions of their leaf shape or flower colour or both.

Table 3 and 4 list some integration results of parallel leaf shape and flower colour descriptions. By comparing the original descriptions and their integrated results, we see that the integration reduces the redundancies in information where possible, for example in the species *Salix pentandra* and *Linum bienne*. On the other hand, if the species itself is variable or authors give widely differing descriptions, the integration process on complementary information retains separate parts of the descriptions, integrating only the parts where authors are in near agreement. The species *Spinacia oleracea* and *Origanum vulgare*) illustrate this, each retaining two separate regions.

6 Querying: combining reasoning with semantic similarity

The representation of continuous quantities in an ontology enables us to carry out semantic-based querying. In this section, we explain how the specially devised reasoner

[4] This is an international project (http://www.efloras.org/) which collects plant taxonomy data from several main floras, such as *Flora of China*, *Flora of North America*, *Flora of Pakistan*, etc. Plant species descriptions are available in electronic form, but written in the usual style of floras: using natural language phrases in a fairly formal framework.

Species	Flower Colour Descriptions	Integration Results		
		R_h	R_s	R_l
Linum	pale blue to lavender			
bienne	pale lilac-blue	63–78	3–80	65–94
(Pale Flax)	pale blue			
	violet-purple			
Origanum	white or purplish-red	0–0	0–0	95–100
vulgare	purple-red to pale pink	80–99	18–88	26–100
(Marjoram)	reddish-purple, rarely white			

Table 4: Examples of integration of parallel flower colour descriptions

interacts with the standard logic reasoner to answer such queries, and how we use similarity ranking to improve the final results.

Consider a search for a species with 'yellow' flowers. If we find a description of flowers 'yellow' (exact match), 'golden' (subsumed by 'yellow'), 'yellow to orange' (subsumes 'yellow') or 'greenish-yellow' (intersects with 'yellow'), then this species is a possible match with the query. These are the four ways of matching which we build into the reasoning.

Following the method used in [WP05, WP06], the query is firstly represented by an OWL-Eu class Q, which equals the species with required leaf shape and/or flower colour. Our reasoner interacts with the FaCT-DG reasoner and returns a list of species whose leaf shapes and/or flower colours satisfy one of the above four ways of matching. Here, automatic reasoning is used to check the relations between species in the knowledge base and the query – the standard qualitative logic reasoning is undertaken. The next stage is to use semantic distance to assess how well a logically-matched species actually matches the query. We use this to rank the results with the most promising matches (i.e. those semantically closest) occurring earliest.

6.1 Query experiments

All queries are based on an integrated knowledge base consisting of 656 species. As described in the previous section, parallel leaf shape and flower colour descriptions were integrated and added into the OWL-Eu ontology. If one species has more than one shape and/or colour region which matches the query, only the *best-match* (that with the smallest distance to the query) is selected for ranking.

Firstly, we carried out 66 queries based on leaf shapes. Each query finished in 1–2 seconds. Table 5 shows the results of a query for a species with 'lanceolate to elliptic' leaves. The matching type indicates the logical relation between the matched species and the query. Incorporating semantic information means that we can find hidden results which are missed by keyword matching, for example, the last species in Table 5. Moreover, matches which would be returned by a keyword search may rightly be ignored by the semantic method in cases where the context of the keyword shifts the denotational region sufficiently far from that of the word itself.

We compared our method with the keyword-based method. For each query, we produced two lists of species: one is returned by our method, the other by keyword matching. For each list, we calculated the distances (using d_{RC}) of all species to the query and returned their average distance as the score of this list. Using the paired sample t-test on these values, we can reject the null hypothesis (that the mean difference

Species	Leaf Shape Descriptions	Matching Type	Distance
Comastoma muliense	lanceolate to elliptic	Exact	0.00
Polygonatum biflorum	narrowly lanceolate to broadly elliptic	Plugin	0.23
Hydrangea longifolia	lanceolate	Subsume	0.85
Rhodiola smithii	linear to oblong narrowly ovate to ovate-linear	Intersection	0.44

Table 5: Some query results for 'lanceolate to elliptic'

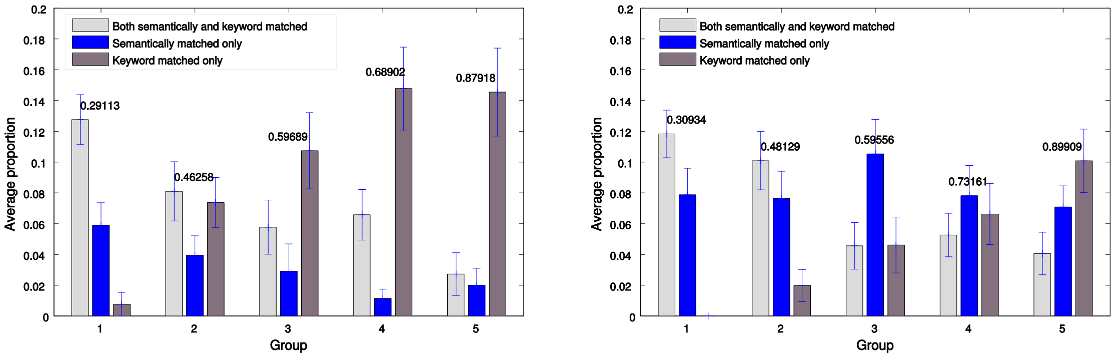
in the scores is zero) at a confidence level greater than 99%. At the same confidence level, the test also shows that the results returned by our method match the queries significantly better.

We have undertaken a further experiment using queries which combine two different continuous quantities. We chose 10 groups of leaf shape and flower colours and used an equal weighting between them. For each query, results returned by our method and keyword matching were reordered into a single list based on their distances to the query. The ordered list was then divided into five groups, representing the top 20% best matched species, 20–40% best matched ones, and so on. In each group, we counted the number of the species that were returned by both methods, those only by our method and those only by keyword matching. The results are displayed in Figure 4 (a). The number above each group of three columns is the average of the mean distance of that group over the 10 queries.

We see that our method is able to find some well-matched results (with small distances) which are not found by keyword matching. However, because of the strictness of logic reasoning (ranges in each dimension have to overlap with those of the query), our method failed to find some results which are actually close to the query and also judged good by experts. Thus, the strict match in each dimension may not always be appropriate. Taking this into account, we decreased the strictness level: if there is at most one dimension not matched for each quantity, the species is still returned if its distance to the query in each quantity is less than the corresponding threshold used for integration. The performance was evaluated similarly, shown in Figure 4 (b). More hidden results were returned by our method while the quality (mean distances) remains stable. A paired sample t-test was also carried out on the combined queries under two different strictness levels (see Table 6). Our method performs significantly better in both situations. We also used precision/recall[5] to measure the query performance of two methods at the different strictness levels (see Table 6). The precision decreases while the recall increases when the strictness of matching becomes loose – a typical balancing problem.

In summary, we have shown that the proposed semantics, when incorporated into methods of integration and querying, enables us to improve considerably on lexical and syntactic approaches.

[5] The precision is the proportion of answers in the returned list which are correct, while the recall is all the correct answers in the dataset that were found. Here, if the distance to the query, in terms of each quantity, is less than the corresponding threshold used for integration, then this species is regarded as matched to the query. The distance function d_{RC} has been validated as described in Section 5.2.



(a) Perfect matching

(b) Partial matching

Figure 4: Comparison between our semantic matching and the keyword matching

Condition	t-Test Value	Semantic Matching		Keyword Matching	
		Precision	Recall	Precision	Recall
Perfectly matched	-4.8966	1.0000	0.1619	0.8718	0.2334
Partially matched	-4.0236	0.9676	0.3121		

Table 6: Comparison between different levels of matching

7 Discussion

The development of this paper is based on (1) modelling continuous quantities, (2) using the general semantics in the special case of city-block metrics in which regions are generalised rectangles, (3) a suitable measure of similarity of regions and, finally, (4) techniques for integration and query. This has proved itself a successful combination, not only in the evaluation but also in its computational tractability, providing us with a semantic basis for information integration and knowledge retrieval.

What about other models and other metrics? There is good evidence that for certain models other metrics are more appropriate, such as the Euclidean metrics (see [RW07]). For other quantities, other metrics may be appropriate according to the differing roles of the dimensions in the modelling space. However, even for Euclidean metrics, regions of the space are, in general, fairly complex polygons (see Figure 2 to see how polygons arise from Euclidean metrics). Distance measures on these regions become correspondingly complex. One of the lessons of this paper is that, in the realm of practical semantics, refinement of models should be balanced against a range of other issues and successful modelling may be achieved with fairly simple models in the context of appropriate semantic processing mechanisms.

The building of an evidential base for semantic processing is itself an interesting undertaking. The expert evidence we have described here to support the semantic analysis is but one form of evidence. Further supporting evidence, for which we have a preliminary analysis [RW07], arises from (1) direct measurements of quantities (for example, leaf shapes and flower colours), and (2) statistical analyses of texts to extract semantic content.

The similarity measure on regions in space that we have presented is of some independent interest, and we are investigating the general construction involved. The whole area of semantic similarity is one of considerable current research.

In this paper we have described continuous quantities using the OWL-Eu ontology language. We plan to investigate its n-ary extension OWL-E [Pan04], which is also supported by the FaCT-DG ontology reasoner. OWL-E is expressive enough to capture more complex modelling spaces, including spaces with dependencies amongst the modelling features.

We have investigated extensively the modelling of only two continuous quantities. Many others appear to fit fairly straightforwardly into this framework. However, interesting issues arise in modelling some kinds of quantities, for example temporal aspects of life histories or spatial aspects of the arrangement of parts. It is clear that much more development is possible in this very practical area of semantics.

7.1 Acknowledgements

The problems tackled in this paper, and its context in the descriptive sciences, arose from Mary McGee Wood's MultiFlora project. David Brée has contributed throughout with valuable discussions on the semantics involved, and discussions with Elzbieta Pekalska on the role of metrics have been useful. Thanks also to the botanical experts who gamely answered our (what must have seemed rather odd) questionnaires.

References

[BBK82] T. Berk, L. Brownston, and A. Kaufman. A new color-naming system for graphics languages. *IEEE Computer Graphics and Applications*, 2(3):37–44, 1982.

[BvHH^{+}04] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein eds. OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/, Feb 2004.

[FDFH90] J. D. Foley, A. Van Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley, 1990.

[GÖ0] P. Gärdenfors. *Conceptual Spaces: the geometry of thought*. The MIT Press, Cambridge, Massachusetts, 2000.

[Lak87] G. Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, 1987.

[LFL98] T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[Lyo95] J. Lyons. *Linguistic Semantics: An Introduction*. C.U.P., 1995.

[MM98] J. R. Massey and J. C. Murphy. Vascular plant systematics: Categorized glossary. http://www.ibiblio.org/botnet/glossary, 1998.

[Moj02] A. Mojsilović. A method for color naming and description of color composition in images. In *Proceedings of the IEEE International Conference on Image Processing, 2002*, pages Vol. 2, 789–792, 2002.

[OST57] C.E. Osgood, G. Suci, and P. Tannenbaum. *The Measurement of Meaning.* University of Illinois Press, Urbana, IL, 1957.

[Pan04] J. Z. Pan. *Description Logics: Reasoning Support for the Semantic Web.* PhD thesis, School of Computer Science, University of Manchester, 2004.

[PH05] J. Z. Pan and I. Horrocks. OWL-Eu: Adding Customised Datatypes into OWL. *Journal of Web Semantics*, 4(1), 2005.

[Res95] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *the 14th International Joint Conference on Artificial Intelligence*, volume 1, pages 448–453, Montreal, 1995.

[RMBB89] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.

[RW07] D. E. Rydeheard and S. Wang. The Manchester Computational Flora website. http://www.cs.manchester.ac.uk/flora, 2007.

[SJ99] S. Santini and R. Jain. Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[Sta97] C. Stace. *New Flora of the British Isles.* Cambridge University Press, 1997.

[SvH04] H. Stuckenschmidt and F. van Harmelen. *Information Sharing on the Semantic Web.* Springer-Verlag, 2004.

[Tom85] S. Tominaga. A color-naming method for computer color vision. In *Proceedings 1985 IEEE Int. Conf. on Cybernetics and Society*, pages 573–577, Tucson, Arizona, 1985.

[Tve77] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[VH01] R. C. Veltkamp and M. Hagedoorn. State-of-the-art in shape matching. In *Principles of Visual Information Retrieval, M. Lew (ed)*, pages 87–119, Springer, 2001.

[WP94] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *the 32th Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, Mexico, 1994.

[WP05] S. Wang and J. Z. Pan. Ontology-based representation and query colour descriptions from botanical documents. In *Proceedings of OTM Confederated International Conferences*, volume 3761 of *Lecture Notes in Computer Science*, pages 1279–1295. Springer, 2005.

[WP06] S. Wang and J. Z. Pan. Integrating and querying parallel leaf shape descriptions. In *Proceedings of International Semantic Web Conference (ISWC2006)*, pages 668–681, Athens, GA, USA, November 2006. Springer.