Semantically Processing Parallel Colour Descriptions(*)

Shenghui Wang[1] and Jeff Z. Pan[2]

[1] School of Computer Science, University of Manchester, UK
[2] Department of Computing Science, University of Aberdeen, UK

Abstract. Information integration and retrieval are useful tasks in many information systems. In these systems, it is far from an easy task to directly integrate information from natural language (NL) sources, because precisely capturing NL semantics is not a trivial issue in the first place. In this paper, we choose the botanical domain to investigate this issue. While most existing systems in this domain support only keyword-based search, this paper introduces an ontology-based approach to process parallel colour descriptions from botanical documents. Based on a semantic model, it takes advantage of ontologies so as to represent the semantics of colour descriptions precisely, to integrate parallel descriptions according to their semantic distances, and to answer colour-related species identification queries. To evaluate this approach, we implement a colour reasoner based on the FaCT-DG Description Logic reasoner, and present some results of our experiments on integrating parallel descriptions and species identification queries. From this highly specialised domain, we learn a set of more general methodological rules.

1 Introduction

Automatic information integration and retrieval have become desirable features for many information systems. The information which these systems have to process is often *descriptive* (written in natural language) and *parallel* (multiple sources describing the same objects or phenomena). Parallel descriptions may emphasise different aspects of the same object; they may represent the same information in different ways, or they may plainly disagree with each other. It is far from an easy task to directly integrate information from natural language (NL) sources, because capturing NL semantics precisely is not a trivial task.

In this paper, we choose the botanical domain to investigate this issue. As one of the premier descriptive sciences, botany offers a wealth of material on which to test our methods. For instance, in our dataset, the species *Origanum vulgare* (Marjoram) has four descriptions of its flowers' colour:

(*) This is an extended and revised version of the paper "Ontology-based Representation and Query of Colour Descriptions from Botanical Documents," which was published in the 4th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE-2005). This work is partially supported by the FP6 Network of Excellence EU project Knowledge Web (IST-2004-507842).

S. Spaccapietra (Ed.): Journal on Data Semantics X, LNCS 4900, pp. 212–236, 2008.

– "violet-purple" in *Flora of the British Isles* [1],
– "reddish-purple, rarely white" in *New Flora of the British Isles* [2],
– "white or purplish-red" in *Flora Europaea* [3],
– "purple-red to pale pink" in *Gray's Manual of Botany* [4].

It has been demonstrated by Wood et. al. [5] that extracting and collecting parallel information from different sources can produce more complete results. Some current projects, such as eFloras[1] and the PLANTS database,[2] attempt to store knowledge from natural language documents in electronic form. These projects generally allow keyword-based queries but fail to process information directly based on their semantics.

This paper makes the following contributions towards semantically processing parallel colour descriptions:

1. It introduces an ontology-based approach to processing parallel colour descriptions from botanical documents. Ideally, an *ontology* captures a shared understanding of certain aspects of a domain: it provides a common *vocabulary*, including important concepts, properties and their definitions, and *constraints* regarding the intended meaning of the vocabulary, sometimes referred to as background assumptions. One of the main advantages of using ontologies is that parallel information can be extracted and represented in a uniform ontology. The explicitly written information can be accessed easily and the implicit knowledge can also be deduced naturally by applying reasoning on the whole ontology. Some earlier work [6,7] has indicated that an ontology could help in extracting, collecting and organising parallel information.
2. It proposes to use a well known colour model, namely the Hue Saturation Lightness (HSL) Model, to model basic colour terms. Based on this semantic model, complex colour descriptions are precisely quantified by applying common morpho-syntactic rules, including adjective modifiers, ranges, conjunctions and disjunctions indicated by NL constructions (see Section 3 for more details). It should be noted that our approach is a general one, and using the HSL model is just one example of a semantic model that can be applied to our approach.
3. It proposes to use the OWL-Eu ontology language [8] to represent the quantitative semantics in the model. OWL-Eu is an extension of the W3C OWL DL [9] standard with unary datatype expressions, which can be used, e.g., to capture the intended quantitative semantics in the HSL Model. The formal representation brings computational and reasoning benefits [10]. For example, subsumption reasoning of the OWL-Eu language can be used to check if one colour description is more general than another one.
4. It presents a framework to support species identification queries. It substantially extends our previous conference paper [11] with the following aspects: (1) For the first time, it proposes two distance functions to calculate distances between parallel information (e.g., the distance between "light blue

[1] http://www.efloras.org
[2] http://plants.usda.gov/

to purple" and "violet-blue to pink"). The first distance function d_1 is based on the hue dimension only, and the second distance function d_3 is based on all three HSL dimensions. The main advantage of these two distance functions is that they are designed for measuring distances between ranges, while existing distance functions can only measure distances between points. These are on the one hand not precise enough to capture the semantic colour model and on the other hand not expressive enough to capture the distance between colour descriptions. (2) Based on the distance functions, an algorithm is provided for integrating parallel colour descriptions. (3) the OWL-Eu subsumption reasoning service can then be used to query the integrated colour descriptions, and the distance functions can be used to rank the answers to such queries.

5. Most importantly, it presents our colour reasoner, which is based on the FaCT-DG DL reasoner, and experiments on species identification queries, including comparing our semantic query with existing keyword-based search. The colour reasoner provides the following functionalities: (1) with the help of a NL parser, it transforms the semantics of colour descriptions into their ontological representations; (2) it collaborates with the FaCT-DG reasoner to answer colour-related species identification queries; (3) it calculates distances of parallel information for integration and also infers some probabilistic conclusions. Furthermore, we present some results of our experiments with the colour reasoner on integrating parallel descriptions and species identification queries (see Section 6 and 8 for more details).

We argue that the ontology-based approach is effective in the colour domain, and we have been investigating its applicability to other domains. We believe that it can also be successfully applied to other domains, as long as an appropriate semantic model is chosen and the domain-dependent aspects are well studied.

The rest of the paper is structured as follows. Section 2 introduces some technical background knowledge of multi-dimensional colour models and the OWL-Eu ontology language. Section 3 presents the morpho-syntactic rules that are used to build complex colour descriptions. Section 4 describes how the semantics of colour descriptions are represented in the OWL-Eu language. Section 5 investigates how to answer species identification queries. Section 6 gives primary experimental results of such queries. Sections 7 and 8 introduce the collaboration of distance measuring and DL reasoning, with some interesting integration results. Some related work is described in Section 9. Finally, Section 10 concludes this paper and discusses some of our future work.

2 Technical Background

2.1 The Colour Model

Several colour representations using a multi-dimensional space (CIE XYZ, L*a*b*, L*u*v*, RGB, CMYK, YIQ, HSV, HSL, etc.) have been employed in computer graphics and image processing. Colours are quantified as points

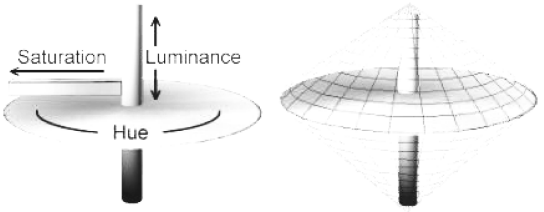

Fig. 1. HSL Colour Model

(or regions) in those spaces. Naming of physically represented colours has been thoroughly investigated [12].

The HSL (Hue Saturation Lightness) model is psychologically based. It corresponds to human's use of colour terms more naturally than machine-oriented colour models, such as the RGB (Red Green Blue) model. In colour notation, it is second only to natural language [13]. The HSL model was therefore chosen to represent basic colour terms. Its colour space is a double cone (see Figure 1).

In the HSL model, a colour is represented by the following three parameters:

- *Hue* is a measure of the colour tint. In fact, it is a circle ranging from 0 (red) to 100 (red again), passing through 16 (yellow), 33 (green), 50 (cyan), 66 (blue) and 83 (magenta).
- *Saturation* is a measure of the amount of colour present. A saturation of 0 is a total absence of colour (i.e. black, grey or white), a saturation of 100 is a pure colour tint.
- *Lightness* (also Luminance or Luminosity) is the brightness of a colour. A lightness of 0 is black, and 100 is white, between 0 and 100 are shades of grey. A lightness of 50 is used to generate a pure colour.

Each basic colour term corresponds to a small space in the double cone whose centre is the particular point representing its HSL value; that is, instead of a point, we represent a colour term by a cuboid space, defined by a range triplet (hueRange, satRange, ligRange). For instance, "purple" is normally defined as the HSL point (83, 50, 25), but is represented by adding a certain range to each parameter, as the region (78–88, 45–55, 20–30).[3]

2.2 OWL DL and Its Datatype Extension OWL-Eu

The OWL Web Ontology Language [15] is a W3C recommendation for expressing ontologies in the Semantic Web. OWL DL is a key sub-language of OWL. Datatype support [16,17] is one of the most useful features that OWL is expected to provide, and has brought extensive discussions in the RDF-Logic mailing list [18] and Semantic Web Best Practices mailing list [19]. Although OWL provides considerable expressive power to the Semantic Web, the OWL datatype

[3] Referring to the NBS/ISCC Color System [14], giving a 100-point hue scale, each major hue is placed at the middle of its 10-point spread, or at division 5.

formalism (or simply *OWL datatyping*) is much too weak for many applications. In particular, OWL datatyping does not provide a general framework for customised datatypes, such as XML Schema user-defined datatypes.

To solve the problem, Pan and Horrocks proposed OWL-Eu [8], a small but necessary extension to OWL DL. OWL-Eu supports customised datatypes through unary datatype expressions (or simply datatype expressions) based on unary datatype groups. OWL-Eu extends OWL DL by extending datatype expressions with OWL data ranges.[4] 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 [8]:

1. *atomic expressions*: if u is a datatype URIref, then $u \in \mathbf{Dexp}(\mathcal{G})$;
2. *relativised negated expressions*: if u is a datatype URIref, 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 $\texttt{or}(E_1, \ldots, E_n) \in \mathbf{Dexp}(\mathcal{G})$.

For example, the following XML Schema user-defined datatype

```
<simpleType name = "HueRange">
  <restriction base = "xsd:integer">
    <minInclusive value = "0"/>
    <maxInclusive value = "100"/>
  </restriction>
</simpleType>
```

can be represented by the following conjunctive datatype expression:

`and`(xsd:nonNegativeInteger, xsdx:integerLessThanOrEqualTo100),

where xsdx:integerLessThanOrEqualTo100 is the URIrefs for the user-defined datatype $\leq_{100}$. Note that *Uniform Resource Identifiers* (URIs) are short strings that identify Web resources [20]. A *URI reference* (or URIref) is a URI, together with an optional fragment identifier at the end. In OWL, URIrefs are used as symbols for classes, properties and datatypes, etc.

Similarly to an OWL DL ontology, an OWL-Eu ontology typically contains a set of class axioms, property axioms and individual axioms. FaCT-DG, a datatype group extension of the FaCT DL reasoner, supports OWL-Eu ontologies.[5] In Section 5, we will use the FaCT-DG reasoner to help answering queries.

3 NL Processing

A close observation of the descriptions in floras shows that colour descriptions are mostly complex phrases, so that they can cover the variations of plant individuals in the field (see the example in Section 1). Complex colour descriptions are built

[4] This is the *only* extension OWL-Eu brings to OWL DL.

[5] To be more precise, FaCT-DG supports the $\mathcal{SHIQ}(\mathcal{G})$ DL, i.e., OWL-Eu without nominals, which are not used in the paper.

Table 1. Colour description patterns and their relative frequencies of occurrence, where X, Y and Z each represent a single colour term or an atomic colour phrase, A is a degree adjective and P is a probability adverb

Description patterns	Frequency of occurrence	Example
X	25.5%	"orange"
A X	36.5%	"pale blue"
X to Y (to Z...)	25.9%	"white to pink to red to purple"
X-Y	19.9%	"rose-pink"
X+ish(-)Y	13.2%	"reddish-purple"
X(, Y) or Z	6.5%	"white or violet"
X(, Y), P Z	6.4%	"reddish-purple, rarely white"
X/Y	4.6%	"pink/white"
X, Y	2.8%	"lavender, white-pink"
X(, Y), and Z	2.3%	"white and green"

from several basic colour terms by applying certain morpho-syntactic rules. In order to be represented correctly, a complex colour description has to be analysed by using the same rules.

We carried out a morpho-syntactic analysis on 227 colour descriptions of 170 species from five floras.[6] Different description patterns and their relative frequencies of occurrence in the data set are summarised in Table 1. Table 3 gives the corresponding BNF syntax for colour descriptions. As shown in Table 1, most patterns describe colour ranges that are built from several atomic colour phrases, such as "blue," "blue-purple" or "bright yellow."

There are two steps in our text processing. Firstly, we construct the following *atomic* colour phrases as basic colour spaces:

X: This is a single colour space, i.e. (hueRange, satRange, ligRange).[7]

A X: We need to modify the space of X according to the meaning of A, as shown in Table 2. For example, "light blue" is represented as (61–71, 70–80, 65–75) where "blue" is (61–71, 90–100, 45–55).

X-Y: This represents an intermediate colour between the two colours X and Y [22]. For example, "blue-purple" is generated as the halfway colour between "blue" (66, 100, 50) and "purple" (83, 50, 25), that is, the colour with HSL value of (75, 75, 38). The hue is calculated by the following formula (with similar calculations for saturation and lightness):

$$Hue_{X-Y} = \frac{Hue_X + Hue_Y}{2} \tag{1}$$

[6] They are *Flora of the British Isles* [1], *Flora Europaea* [3], *The New Britton and Brown Illustrated Flora of the Northeastern United States and Adjacent Canada* [4], *New Flora of the British Isles* [2] and *Gray's Manual of Botany* [21].

[7] According to the Colour Naming System (CNS) [22], given a 100-point hue scale, each major Munsell hue placed at the middle of its 10-point spread, or at division 5. Therefore, for each basic term, a 5-point spread along each side of the prototypical values builds up a reasonable space. This setting is inherited by some of the following operations.

Table 2. Meanings of modifiers and their corresponding operations on a colour space

Adjective	Meaning[a]	Operation[b]
strong	high in chroma	satRange + 20
pale	deficient in chroma	satRange - 20, ligRange + 20
bright	of high saturation or brilliance	satRange + 20, ligRange + 20
deep	high in saturation and low in lightness	satRange + 20, ligRange - 20
dull	low in saturation and low in lightness	satRange - 20, ligRange - 20
light	medium in saturation and high in lightness	satRange - 20, ligRange + 20
dark	of low or very low lightness	ligRange - 20

[a] Referring to Merriam-Webster online dictionary.
[b] Referring to the specifications from the Colour Naming System (CNS) [22], saturation and lightness are each divided into 5 levels, which causes a range/ranges to change by 20 (100/5).

Finally it is represented by the range triple (70–80, 70–80, 33–43), by adding 5-point spread in each dimension from the centre.

Xish-Y: Specified in CNS [22], this denotes a quarterway value between the two colours, closer to the latter colour term. For instance, "reddish-purple" means it is basically purple (83, 50, 25) but reflecting a quarterway deviation to red (100, 100, 50), so the hue range for "reddish-purple" is centred on 87, calculated by the following formula (similar formulae for saturation and lightness):

$$Hue_{X_{ish}-Y} = Hue_Y + \frac{Hue_X - Hue_Y}{4} \tag{2}$$

and the colour is finally represented as (82–92, 58–68, 29–39).

Secondly, we build up combined colour spaces based on basic ones. Specifically, combined colour spaces are built up by a colour reasoner, according to the following morpho-syntactic rules:

1. If atomic colour phrases are connected by one or more "to"s, the final colour space should be the whole range from the first colour to the last one. For instance, if "light blue" is (66, 80, 70) and "purple" is (83, 50, 25), "light blue to purple" should be the whole range (66–83, 50–80, 25–70), which contains any colour in between.

   Note that special care is needed for ranges starting or ending with a grey colour, such as "white to purple." In the HSL model, colours ranging from white, through different levels of grey, to black have no hue and saturation values. For instance, the HSL value of "white" is (0, 0, 100), while "red" also has a hue value of 0 but its saturation is 100. A special rule for building such ranges has to be followed; that is, a range from colour A (0, 0, l_a) to colour B (h_b, s_b, l_b) should be ($\overline{h_b - 5}$–$\overline{h_b + 5}$, 0–s_b, l_a–l_b), where the hue value does not range from 0 to h_b which is actually from red to colour B. For example, the range from "purple" (83, 50, 25) to "white" (0, 0, 100) should

Table 3. BNF syntax of colour descriptions

$< Cterm >$	$::= red \mid yellow \mid green \mid \ldots$
$< Dmodifier >$	$::= strong \mid pale \mid bright \mid deep \mid dull \mid light \mid dark \mid \ldots$
$< Pmodifier >$	$::= usually \mid often \mid sometimes \mid occasionally \mid rarely \mid never \mid \ldots$
$< Cphrase >$	$::= < Cterm >$
	$\mid < Cterm > [ish][- \mid \ ] < Cterm >$
	$\mid < Cphrase > - < Cphrase >$
	$\mid < Dmodifier > < Cterm >$
$< Cdescription >$	$::= < Cphrase >$
	$\mid < Cphrase > \{\ to\ < Cphrase >\}$
	$\mid < Cphrase >,\ < Cphrase >$
	$\mid < Cphrase > / < Cphrase >$
	$\mid < Cphrase > \{, < Cphrase >\}\ or\ < Cphrase >$
	$\mid < Cphrase > \{, < Cphrase >\}\ and\ < Cphrase >$
	$\mid < Cphrase > \{, < Cphrase >\},\ < Pmodifier > < Cphrase >$

be represented by the triple (78–88, 0–50, 25–100), so that the hue range (78–88) keeps the purple tint when the colour changes from purple to white.

2. If atomic colour phrases are connected by any of these symbols: "or," "and," comma (",") or slash ("/"), they are treated as separate colour spaces; that is, they are disjoint from each other. For instance, "white, lilac or yellow" means that the colour of this flower could be either white or lilac or yellow, not a colour in between.

   Notice that "and" is treated as a disjunction symbol because, in floras, it normally means several colours can be found in the same species, instead of indicating a normal logical conjunction. For instance, flowers of species *Rumex crispus* (Curled Dock) are described as "red and green," which means that both red and green flowers may occur in the same species, but it does not mean that one colour is both red and green.

By using an NL parser based on our BNF syntax, we can generate an OWL-Eu ontology to model complex colour information.

4 Representation of Colour Descriptions in OWL-Eu

Based on the morpho-syntactic rules introduced in the last section, we can decompose the semantics of colour descriptions into several quantifiable components, which can be represented as DL datatype expressions. In this section, we will show how to use the OWL-Eu ontology language to represent the semantics of a colour description.

The fragment of our plant ontology $\mathcal{O}_C$ contains Colour as a primitive class. Important primitive classes in $\mathcal{O}_C$ include

```
Class(Species), Class(Flower), Class(Colour);
```

important object properties in $\mathcal{O}_C$ include

```
ObjectProperty(hasPart), ObjectProperty(hasColour);
```

important datatype properties in $\mathcal{O}_C$ include

```
DatatypeProperty(hasHue Functional
 range(and(xsd:nonNegativeInteger, xsdx:integerLessThanOrEqualTo100))),
DatatypeProperty(hasSaturation Functional
 range(and(xsd:nonNegativeInteger, xsdx:integerLessThanOrEqualTo100))),
DatatypeProperty(hasLightness Functional
 range(and(xsd:nonNegativeInteger, xsdx:integerLessThanOrEqualTo100))),
```

which are all functional properties. A *functional* datatype property relates an object with at most one data value. Note that the datatype expression

```
and(xsd:nonNegativeInteger, xsdx:integerLessThanOrEqualTo100)
```

is used as the range of the above datatype properties.

Based on the above primitive classes and properties, we can define specific colours, such as Purple, as OWL-Eu defined classes (indicated by the keyword "complete") .

```
Class(Purple complete Colour
  restriction(hasHue someValuesFrom
                (and(xsdx:integerGreaterThanOrEqualTo78,
                 xsdx:integerLessThanOrEqualTo88)))
  restriction(hasSaturation someValuesFrom
                (and(xsdx:integerGreaterThanOrEqualTo47,
                 xsdx:integerLessThanOrEqualTo52)))
  restriction(hasLightness someValuesFrom
                (and(xsdx:integerGreaterThanOrEqualTo20,
                 xsdx:integerLessThanOrEqualTo30))))
```

In the above class definition, datatype expressions are used to restrict the values of the datatype properties *hasHue*, *hasSaturation* and *hasLightness*. Note that not only colour terms but complex colour descriptions can be also represented in OWL-Eu classes, as long as they can be transformed into proper colour subspaces with constraints on their hue, saturation and lightness.

As colour descriptions are represented by OWL-Eu classes, we can use the subsumption checking service provided by the FaCT-DG reasoner to check if one colour description is more general than another. Namely, if ColourA is subsumed by ColourB, we say that ColourB is more general than ColourA. With the help of the FaCT-DG DL reasoner, the formal representation of colour descriptions makes it possible to express a query about a range of colours, such as to retrieve all species which have "bright rose-pink" or "light blue to purple" flowers.

5 Domain-Oriented Queries

The flower colour of an individual plant is an important distinguishing feature for identifying which species it belongs to. The species identification that botanists are interested in can be written as a query: "Given a certain colour, tell me all

the possible species whose flowers have such a colour." We would like to point out that, from a botanical point of view, one has to take the variations between individuals in nature into account. In other words, botanists rarely use colour as a strict criterion. It is more appropriate to answer such species identification queries in an fuzzy manner, that is, returning a list which contains all species that *could* match the query. We call this kind of query, which is particularly suitable for domain interests, *domain-oriented* queries.

We can answer species identification queries based on subsumption queries that are supported by the FaCT-DG DL reasoner. For example, if the plant ontology contains the following class axioms:

```
Class(SpeciesA restriction(hasPart someValueFrom(FlowerA)))
Class(FlowerA restriction(hasColour someValueFrom(ColourA)))
Class(SpeciesB restriction(hasPart someValueFrom(FlowerB))
Class(FlowerB restriction(hasColour someValueFrom(ColourB)))
```

and if from the definitions of ColourA and ColourB we can conclude that ColourA is subsumed by ColourB, when we ask our DL reasoner whether the above ontology entails that SpeciesA is subsumed by SpeciesB, the reasoner will return "yes." By using this kind of subsumption query, we can, for example, conclude that a species having "golden" flowers is subsumed by a more general species which has "yellow" flowers, which again is subsumed by another species which has "orange to yellow" flowers. Therefore, if one asks "Which species might have yellow flowers," our colour reasoner will return all these three species.

For species identification, this hierarchical subsumption matching is very useful for shortening the possible species list. After classification reasoning, we have already had three different levels of matchings:

- **Exact** matching ($\mathsf{Class}_{\mathsf{RealSpecies}} \equiv \mathsf{Class}_{\mathsf{QuerySpecies}}$),
- **PlugIn** matching ($\mathsf{Class}_{\mathsf{RealSpecies}} \sqsubseteq \mathsf{Class}_{\mathsf{QuerySpecies}}$)
- **Subsume** matching ($\mathsf{Class}_{\mathsf{RealSpecies}} \sqsupseteq \mathsf{Class}_{\mathsf{QuerySpecies}}$)

Actually there is another possible species list, which is not covered by the above three kinds of matchings, that is, **Intersecting** matching ($\neg(\mathsf{Class}_{\mathsf{RealSpecies}} \sqcap \mathsf{Class}_{\mathsf{QuerySpecies}} \sqsubseteq \bot)$) [23,24]. For example, if a species has "greenish-yellow" flowers, it would also be possible to find in the field an individual which has "yellow" flowers. Although this latter list has a lower probability to contain the correct answers, it is still helpful from botanical point of view.

Our colour reasoner reduces our domain problems into standard DLs reasoning problems. In fact, in order to answer domain-oriented queries, it interacts with the FaCT-DG reasoner. First, the colour in a query is represented by an OWL-Eu class Q with datatype constraints about its hue, saturation and lightness.

Secondly, the colour reasoner calculates the complete set of colours $complete_Q$ which satisfies the above four levels of matching. Specifically, $complete_Q$ consists of the following four sets.

- $equiv_Q$: all elements are equivalent to the class Q, such as "yellow;"
- sub_Q: all elements are subsumed by the class Q, such as "golden;"

- $super_Q$: all elements subsume the class Q, such as "yellow to orange to red;"
- $intersection_Q$: all elements intersect with the class Q, such as "greenish-yellow."

Note that the first two contain answers with 100% confidence, while the latter two contain those with less confidence. Thirdly, in order to find all species that have flowers whose colour satisfies the query, the colour reasoner interacts with the Fact-DG reasoner to return those species which have flowers whose colour is contained in $complete_Q$ set.

6 Experiments on Representation and Query

In this section, we will present some experiments, based on our plant ontology, of species identification queries.

We chose 100 colour terms which are commonly found in floras, as basic colour terms. For each basic term, we obtained its RGB value by referring to the X11 Colour Names,[8] converted this into its corresponding HSL value and finally defined it as ranges in hue, saturation and lightness (as described in Section 4).

A simple plant ontology, mentioned in Section 4, was constructed using the OWL-Eu language. This ontology contains 1154 species, selected from five floras, mentioned before, and the online eFloras.[9], each of which has a flower part which has a colour property. The colour property is represented by a datatype expression, representing the colour spaces transformed from the original colour descriptions,

For example, species *Viola adunca* has "light blue to purple" flowers.

```
Class(Viola_adunca complete Species
  restriction(hasPart someValuesFrom(Viola_adunca_flower))),
Class(Viola_adunca_flower complete Flower
  restriction(hasColour someValuesFrom(Viola_adunca_flower_colour))),
Class(Viola_adunca_flower_colour complete Colour
  restriction(hasHue someValuesFrom
                (and(xsdx:integerGreaterThanOrEqualTo66,
                 xsdx:integerLessThanOrEqualTo83)))
  restriction(hasSaturation someValuesFrom
                (and(xsdx:integerGreaterThanOrEqualTo50,
                 xsdx:integerLessThanOrEqualTo100)))
  restriction(hasLightness someValuesFrom
                (and(xsdx:integerGreaterThanOrEqualTo25,
                 xsdx:integerLessThanOrEqualTo70))))
```

In our experiments, 10 species identification queries based flower colours were

[8] http://en.wikipedia.org/wiki/X11_Color_Names

[9] This is an international project 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 still written in the common style of floras, i.e. semi-NL.

Table 4. Query results (partial) of species having "yellow," "light blue" and "light blue to purple" flowers

Species	Flower colour	Matching type
Amsinckia menziessi	yellow	Exact matching
Ranunculus acris	golden	PlugIn matching
Eucalyptus globulus	creamy-white to yellow	Subsume matching
Tropaeolum majus	yellow to orange to red	Subsume matching
Rhodiola sherriffii	greenish-yellow	Intersection matching
Eschscholzia californica	deep orange to pale yellow	Intersection matching

(a) "yellow"

Species	Flower colour	Matching type
Aster chilensis	light blue	Exact matching
Heliotropium curassavicum	white to bluish	Subsume matching
Linum bienne	pale blue to lavender	Subsume matching
Triteleia laxa	blue to violet	Intersection matching
Dichelostemma congestum	pink to blue	Intersection matching

(b) "light blue"

Species	Flower colour	Matching type
Viola adunca	light blue to purple	Exact matching
Linum bienne	pale blue to lavender	PlugIn matching
Verbena lasiostachys	blue-purple	PlugIn matching
Lupinus eximus	blue to purple, sometimes lavender	Intersection matching
Stachys bullata	light purple to pink to white	Intersection matching
Triteleia laxa	blue to violet	Intersection matching

(c) "light blue to purple"

carried out. The queries consist of basic terms, range phrases and others with different levels of complexity (as shown in Table 1). Each query finished in 1–2 seconds on a 2G Hz Pentium 4 PC. Some of the results are presented in Tables 4, in the order of complexity of colours: "yellow," "light blue," "light blue to purple."

We can query in a specific manner, for example to find species which have "light blue" flowers but excluding those with "dark blue" flowers (see Table 4 (b)); or in a more general style, to query all species which could have flowers ranging from "light blue to purple" (see Table 4 (c)). All of these facilities use our quantitative model which makes it possible to compare and reason with classes at a semantic level.

As stated in Section 5, the resulting list is from four different levels of matching, which gives a complete list for species identification. We can also specify to stop at certain levels of matching to get results with different confidences, such as only returning those species which fully satisfy the query.

The semantics of a colour term or a complex colour description is decomposed and represented by a group of ranges in multiple numerical parameters, which is a small subspace in a multi-dimensional space. Numerical representation makes

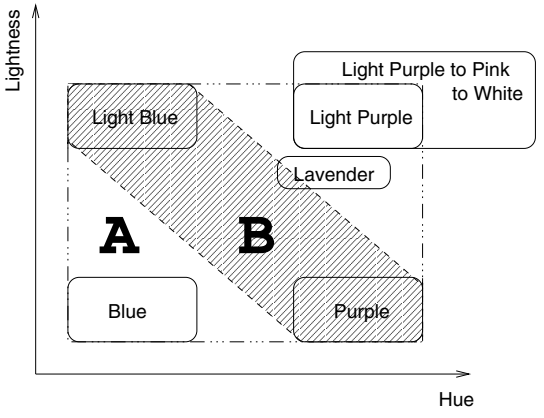

Fig. 2. Range between "light blue" and"purple"

Table 5. Performance comparison between semantic matching and keyword matching

Method	Precision	Recall
Semantic matching	98.2%	81.1%
Keyword matching	84.8%	71.9%

it easy to build ranges between colours, but a further observation shows that this is not as obvious as we thought. For example, there could be different ways of interpreting the meaning of "light blue *to* purple" (see Figure 2):

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

In our experiment (see Table 4 (c)), we used the last option (the whole rectangle) for the sake of simplicity and computation cost. It is open to extend our work and to allow the users to pick up one of the above options when they query with the keyword "to."

We further compared our semantics-based query with the simple keyword matching. The standard precision and recall[10] were use to measure the performance. Here, if the distance of returned answer to the query is less than a threshold used for integration, as we will introduce later, then this answer is considered correct. Table 5 gives the comparison results of the performance of these two methods.

[10] The precision indicates the proportion of answers in the returned list which are correct, while the recall is all the correct answers in the whole dataset that were found.

7 Integration of Parallel Colour Descriptions

In the previous sections, we have shown that, by using a multi-dimensional colour model, we can precisely represent the semantics of complex colour descriptions. Represented in the OWL-Eu language, this quantitative representation enables reasoning on the real semantics of NL information and provides more practical query results.

However, the reality does not stop here. As the example in Section 1 shows, a species is often observed by different botanists, that is, parallel descriptions of the same species are easily found cross different floras. Since they describe the same species, these parallel colour descriptions are expected to be similar to or compatible with each other. Most importantly, as demonstrated in [5], extracting and collecting parallel information from different sources can produce more complete results.

A close observation of real data shows that, even using a standard naming system, botanists use their personally preferred patterns to describe what they observe, so as to cover the variations of plant individuals in nature. Accordingly, parallel colour descriptions are rarely exactly the same; sometimes they vary a lot, especially when the species itself has a relatively wide variation. Here we do not focus on some genuine geographic or temporal influences on species variation or some literary errors; instead, we are only concerned linguistic differences between parallel information. We assume that information from different sources is correct but probably incomplete; i.e. different sources are never considered contradictory, only complementary, possibly with a certain degree of overlap.

The key task is to find good strategies to integrate parallel colour descriptions; otherwise, we could end up producing incomplete or redundant results. A simple *conjunction* (or *intersection*) would cause information loss. For example, if one flora says that a flower is "white" while another says it is "white or purplish, sometimes yellow," the result of their intersection is "white"—"purple" and "yellow" would be removed. Another logic operation, *disjunction* (or *union*), does not work ideally either. For example, there are two descriptions of the same flower: "reddish-purple, rarely white" and "white or purplish-red." Each of them is represented by two separate colour subspaces. The union operation results in

$$\mathsf{White} \sqcup \mathsf{Purplish\text{-}red} \sqcup \mathsf{Reddish\text{-}purple}$$

Note that the other "white" is omitted because two "white"s are identical. The result is complete but there is a redundant overlap between "reddish-purple" and "purplish-red;" while people can easily infer from the original texts that actually any colour between red and purple is possible for this species.

The above observations show that naive use of logic operations cannot produce the integration results as we really expect, indicating that it is not appropriate to simply mix information without careful studies of how similar or how different they are. Along this line, investigations of the similarities of parallel information seem to be a good integration strategy. To a large degree, similarities of their semantics can tell how much different descriptions agree with each other,

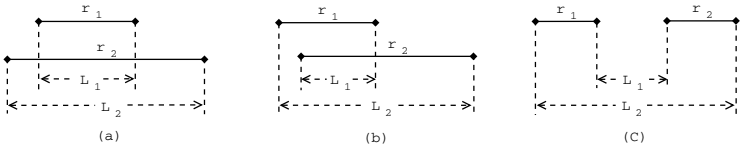

Fig. 3. Three different relations between two ranges

namely, the more similar two descriptions are, the more compatible they are. Thus, given similarities are quantified properly, if two descriptions are similar enough, although they might not be exactly the same (due to various reasons), it is better to combine them as one single "super-description" and remove redundancies; otherwise, it is safer to leave them separate because they are both likely to provide partial information of the same object.

The similarity of two objects is often closely related to the distance between their representations in certain underlying spaces [25]. More specifically, similarity is a decaying function of distance. Ideally, since colour is a common perceptual phenomenon, any distance function for colours should be able to capture the real differences perceived by human eyes. However, how to find perfect colour distances in different colour models is beyond the scope of this paper. Here, we claim that any perceptually acceptable distance function $d(x, y)$ (for a metric space $\mathcal{S}$) which satisfies the following conditions (for all points x, y in $\mathcal{S}$) will suffice.

Minimality: $d(x, y) = 0 \implies x = y$;
Symmetry: $d(x, y) = d(y, x)$.

In what follows, we will present two ways to define the distance function $d(x, y)$; we will also show that both distances satisfy the above two conditions.

Inspired by Tversky's feature contrast and ratio model [26], given two ranges r_1 and r_2, the distance of r_1 and r_2, i.e. $d(r_1, r_2)$, is equal to the "non-common part of r_1 and r_2" divided by "the minimal super-range that contains both r_1 and r_2".

Distance Function $d1$**.** We start to consider a simple distance function: distance w.r.t. the hue-range only. Obviously, hue differences is always the first and the most prominent aspect when people try to compare colours.

There are 3 different types of relations between two ranges, shown in Figure 3. With the help of the FaCT-DG DL reasoner, we can tell whether one range subsumes the other ($r_1 \sqsubseteq r_2$), or they intersect with each other ($\neg(r_1 \sqcap r_2) \sqsubseteq \bot$), or they are disjoint from each other ($(r_1 \sqcap r_2) \sqsubseteq \bot$). Accordingly, we define the following distance function for two arbitrary ranges r_1 and r_2:

$$d1(r_1, r_2) = \begin{cases} 1 - \frac{L_1}{L_2} & \text{if } r_1 \text{ and } r_2 \text{ overlap} \\ 1 + \frac{L_1}{L_2} & \text{otherwise;} \end{cases} \tag{3}$$

where L_2 is the length of minimal super-range which contains both 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 part of two ranges; otherwise, for (c), L_1 is the length of the gap between two ranges. If two ranges r_1 and r_2 only share one point, we say they *meet* each other and $L_1 = 0$

The distance $d1(r_1, r_2)$ is continuous and nicely scaled into the range $[0, 2)$: if $d1(r_1, r_2) = 0$, r_1 equals r_2; if $0 < d1(r_1, r_2) < 1$, r_1 and r_2 overlap; if $d1(r_1, r_2) = 1$, r_1 meets r_2; if $1 < d1(r_1, r_2) < 2$, r_1 and r_2 are disjoint; as two ranges move further apart from each other, the distance gets closer to 2.

Distance Function $d3$**.** As we know, hue, saturation and lightness values should be assigned to a colour at the same time because they are *integral* dimensions [27]. In order to have a more sensible distance measure, it might be better to take the other two dimensions into account.

We still use the overlap/gap ratio to measure distances. Instead of comparing the length of ranges in one dimension, we measure the volume of the overlap/gap space. Similarly, the FaCT-DG DL reasoner helps to classify the relation between two colour spaces, which would be subsumption, intersection or disjunction. Accordingly, we define the function $d3$ for two colour spaces cs_1 and cs_2:[11]

$$d3(cs_1, cs_2) = \begin{cases} 1 - \frac{V_1}{V_2} & \text{if } cs_1 \text{ and } cs_2 \text{ overlap} \\ 1 + \frac{V_1}{V_2} & \text{otherwise} \end{cases} \tag{4}$$

where V_2 is the volume of minimal cuboid space which contains both cs_1 and cs_2, and V_1 is defined as follows: when cs_1 and cs_2 overlap with each other, V_1 is the volume of the overlap space of the two spaces; otherwise, V_1 is the volume of the gap between two spaces in terms of their "super-space" V_2. It is easy to show that the distance function 4 has exactly the same properties as the distance function 3 has.

Once the distances of any two colour descriptions are calculated, users can have a better overview of all parallel information from different sources. Based on such an overview, they can therefore decide whether it is necessary to combine two pieces of information or just to leave them as separate as they are. If a reasonable distance threshold is given,[12] our colour reasoner automatically combines two descriptions if they are close/similar enough or keeps them separate otherwise.

The integration process is recursive as follows:

Step 1. Use the FaCT-DG DL reasoner to classify the relations between any two colour spaces generated from parallel descriptions of the same species, and then use our colour reasoner to calculate their distances (by using either Formula 3 or 4).

Step 2. Select two closest colour spaces and check whether they are "similar-enough," i.e. their distance is less than the distance threshold.

[11] Here, $d3$ means the distance function considers all three dimensions, instead of only the hue is considered as $d1$ measures.

[12] See Section 8 for more detail.

Step 3. If they are not similar enough then the integration stops; otherwise, the smallest cuboid space which contains them is generated and substitutes them as their integrated space (the same operation as building "to" ranges in Section 3).
Step 4. Go back to Step 1 to check the updated colour spaces.

In the final results, not only the integrated colour spaces are stored, but also those generated from parallel sources are kept for further references. For each disjoint colour space rcs_i in the final results, we check how many of the original colour spaces intersect with it.

$$Prob_{rcs_i} = \frac{\text{Count of original colour spaces that intersect with } rcs_i}{\text{Count of original colour spaces}} \tag{5}$$

According to the $Prob$ value of each original colour space, we can see how many authors agree on one particular range of colours, which reflects how likely people will find such coloured plant individuals in the field. Therefore, some interesting frequency inferences can be deduced from parallel information integration, which will be illustrated in the next section.

8 Experiments on Integration

In this section, we present some results of our experiments on the integration of parallel colour descriptions. These experiments illustrate how the collaboration of DL reasoning and similarity measuring helps to integrate parallel information. Interestingly, our results can also be used to evaluate the performance of the two similarity functions in a real application.

We further selected 656 species, each of which has at least two parallel descriptions. Note that due to geographic influences, i.e. some species only exist in some particular regions, parallel information is not guaranteed for each species.

We extended the NL parser introduced in Section 3 in order that it can parse a whole botanical document and extract flower colour descriptions before it deeply parses these colour descriptions by using morpho-syntactic rules (see Table 3). All data is extracted by the parser automatically and double-checked manually.

In order to calculate the threshold for the integration, we selected a group of parallel descriptions from the whole dataset, which are not identical yet are still considered to be similar enough to be combined. The average distance of these parallel descriptions is used as the threshold. Interestingly, we got slightly different thresholds for two similarity functions, i.e. 1.5 for $d1$ and 1.4 for $d3$.

To simplify the presentation, here we use two species to illustrate our experiments. According to three different authors, *Linum bienne* (Pale Flax) has "pale blue to lavender," or "pale lilac-blue" or "pale blue" flowers. In the 3D HSL-space, the FaCT-DG DL reasoner classifies their relations as follows:

- $\neg(CS_{pale\ blue\ to\ lavender} \sqcap CS_{pale\ blue}) \sqsubseteq \bot$ ("pale blue to lavender" intersects with "pale blue"),
- $(CS_{pale\ blue\ to\ lavender} \sqcap CS_{pale\ lilac-blue}) \sqsubseteq \bot$ ("pale blue to lavender" is disjoint from "pale lilac-blue"), and

– $(CS_{pale\ blue} \sqcap CS_{pale\ lilac-blue}) \sqsubseteq \perp$ ("pale blue" is disjoint from "pale lilac-blue").

According to their logic relations, their distances are calculated differently. By using function $d3$, distances between these colour spaces are:

– $d3(CS_{pale\ blue\ to\ lavender}, CS_{pale\ blue}) = 0.55$,
– $d3(CS_{pale\ blue\ to\ lavender}, CS_{pale\ lilac-blue}) = 1.26$,
– $d3(CS_{pale\ blue}, CS_{pale\ lilac-blue}) = 1.77$.

$CS_{pale\ blue\ to\ lavender}$ and $CS_{pale\ blue}$ are combined first because they are close enough (actually, $CS_{pale\ blue} \sqsubseteq CS_{pale\ blue\ to\ lavender}$, so $CS_{pale\ blue\ to\ lavender}$ is kept as their integrated space), then the integration process goes back to check the newly updated colour spaces. This time, $CS_{pale\ lilac-blue}$ is close enough (1.26 is less than the threshold for $d3$, which is 1.4) to the newly integrated colour space ($CS_{pale\ blue\ to\ lavender}$), they are combined too although they do not overlap with each other directly. Therefore, three slightly different NL descriptions are finally combined as one single and unified colour space.

Differently, the species *Allium dichlamydeum* (Coast Onion) has two descriptions about its flower colour: "pink to rose" and "deep reddish-purple". They are obviously disjoint from each other; their distance is 1.63 which is higher than the threshold, so they are kept separately.[13] Table 7 shows more examples of parallel data and their integration results.

Our experiments confirm that the different effects of two distance functions $d1$ (based on hue dimension only) and $d3$ (based on all three HSL dimensions). Again taking *Allium dichlamydeum* (Coast Onion) for example, if only the hue dimension is considered, the two descriptions would be combined as a single colour space because their hue ranges are actually quite similar. However, after taking saturation and lightness into account, the HSL-space similarity function successfully keeps them separate, which seems more acceptable to human perception. Other similar cases are shown in Table 6.

It might be expected that using all three HSL dimensions would lead to very different integration results to those using the distances in the single hue dimension. Interestingly, these two distance functions give almost the same results in most cases. Only 20% of the parallel data give different results; for example, in Table 7, both distance functions (with different thresholds) give exactly the same integrated results. The more complicated HSL-space distance function ($d3$) does not produce as much advantage as we had expected. One possible reason is, as we mentioned in Section 7, that although people use different modifiers to distinguish colours' saturation and lightness, hue is still the most prominent aspect which really counts for describing flower colours. Therefore we choose to use the simpler hue-range distance as the default criterion for integration, while HSL-space distance is used for some advanced comparisons.

As stated in Section 7, one of the advantages of processing parallel information is that we can infer some probabilistic conclusions by observing how often certain information is mentioned by different authors, as the last column in

[13] It has been checked out that this species has slightly different flower colour according to its geographic distribution.

Table 6. Comparison of integration results from two different distance functions

Species	Parallel Descriptions	Integration Results			
		Distance Function	H	S	L
Allium dichlamydeum	pink to rose deep reddish-purple	$d1$	84–0	13–61	6–87
		$d3$	84–90	22–61	6–16
			97–0	13–24	45–87
Iris laevigata	blue dark blue or violet	$d1$	63–86	39–100	25–77
		$d3$	80–86	39–45	67–77
			63–69	60–100	25–55
Hylotelephium ewersii	pink or light purple purplish-red	$d1$	80–0	22–50	20–92
		$d3$	84–0	22–26	82–92
			80–86	35–50	20–50

Table 7 shows. Looking back to the example mentioned in Section 1, flowers of *Origanum vulgare* (marjoram) have been described by four different authors. After integration, "violet-purple," "purplish-red," "purple-red to pale pink" and "reddish-purple" are combined and substituted by the colour space whose hue ranges from 80 to 99, saturation from 18 to 88 and lightness from 26 to 100; "white" is kept as a disjoint colour space found from parallel sources. The former colour space has a higher probability value (66.7%) than the latter one (33.3%), from which a reasonable inference can be deduced that *white* marjoram flowers are less likely to be found in nature.

Table 7. Examples of parallel descriptions and their integration results

Species	Parallel Descriptions	Integration Results			
		H	S	L	Prob
Lathyrus latifolius	bright rose-pink vivid magenta-pink rose-pink	87–2	13–50	61–91	100%
Linum bienne	pale blue to lavender pale lilac-blue pale blue	63–78	3–80	65–94	100%
Raphanus sativus	lavender, white-pink	63–0	5–50	20–99	66.7%
	white or violet	0–0	0–0	95–100	22.2%
	white, lilac or violet, rarely purple/yellow	13–19	95–100	45–55	11.1%
Ranunculus arvensis	lemon-yellow pale greenish-yellow	12–23	60–96	46–75	100%
Origanum vulgare	violet-purple white or purplish-red	80–99	18–88	26–100	66.7%
	purple-red to pale pink reddish-purple, rarely white	0–0	0–0	95–100	33.3%

9 Related Work

Automatically integrating information from a variety of sources has become a necessary feature for many information systems [10]. Compared to structured or semi-structured data sources, information in natural language documents is more cumbersome to access [28]. Our work focuses mainly on parallel information extraction and integration from homogeneous monolingual (English) botanical documents.

Information Extraction (IE) [29] is a common Natural Language Processing (NLP) technique which can extract information or knowledge from documents. Ontologies, containing various semantics expressions of domain knowledge, have recently been adopted in many IE systems [30,31,32]. Semantics embedded in ontologies can boost the performance of IE in terms of precision and recall [33]. Since they can be shared by different sources, ontologies also play an important role in the area of information integration [10,34,28]. Ontology reasoning is also introduced into the extraction, representation and integration processes [35,36,33]. We have shown that reasoning support for ontologies with customised datatypes is very useful for answering species identification queries and integration of parallel colour descriptions.

One of our main contributions is to capture the NL semantics as precisely as possible. In other research areas, many methods have been tried to solve similar problems. Semantic differential [37] measures people's reactions to words or concepts in terms of ratings on bipolar scales defined with contrasting adjectives at each end, such as "good–bad". Individuals' connotations are captured in a multidimensional space and thus the psychological "distance" between words or concepts are measured. Lexical Decomposition [38] attempts to break the meanings of words down to several basic categories, hoping to find some internal structure for words' meaning. Multidimensional modelling was also employed in several areas of cognitive science [25]. Spatial or geometrical structures are exploited in concept formation and learning, and also in studies in cognitive linguistics [39]. The limitations of their methods are either the dimensions are difficult to interpret or they are most qualitative which prevents to capture semantics precisely.

The quantitative semantic model can produce more useful results for real domain purposes. Specifically, in the botanical domain, many current plant databases can only support keyword-based query, such as the ActKey,[14] ePIC project,[15], the PLANTS database,[16] etc. They rely heavily on the occurrence of keywords. As demonstrated in Section 6, our method uses real semantics matching, instead of pure keyword matching, which supports more flexible-styled queries, especially range-based ones.

Another important related research area is semantic similarity measurement. Obviously, similarity is an important criterion for integration. Depending on how

[14] http://flora.huh.harvard.edu:8080/actkey/
[15] http://www.rbgkew.org.uk/epic/
[16] http://plants.usda.gov/

they are represented in different models, similarity between objects is calculated differently, such as the ratio of common/distinct features in *feature models* [26], the vector distances in multidimensional *spacial models* [25,40], the path-length in *network models* [41,42], etc. In NL research, corpus-based methods are often used to measure similarities between concepts by comparing their information content [43]. Unfortunately, these methods only focus on relations between basic terms, but rarely pay enough attention to more complex expressions, such as regions or ranges. In other words, they are probably able to find the similarity between "lilac" and "purple," but cannot tell how close "lilac to pale blue" is to "deep reddish-purple," which is much more common in the real world. Our method uses a 3D-space as a basic representation of basic colour terms and maps all common linguistics rules into operations on such spaces. Complex NL descriptions are represented by one or several subspaces. By calculating the distances between these subspaces, the similarities between their original NL descriptions are successfully quantified and therefore used as a crucial criterion for the integration.

10 Conclusion and Outlook

This paper has presented and evaluated an ontology-based approach which facilitates representing, integrating and querying colour information from parallel floras. It turns out that, even in this limited domain, formally representing the semantics of colour descriptions is not a trivial problem. Based on a multidimensional semantic model and certain morpho-syntactic rules, we have implemented an NL parser which translates complex colour descriptions into quantitative representations written in the OWL-Eu ontology language. A colour reasoner is implemented to interact with the FaCT-DG DL reasoner in order to integrate parallel information and carry out queries for real botanical applications.

We have shown that our approach outperforms keyword-based approaches, which are widely used in this domain. Firstly, our quantifiable model enables automatic reasoning on the real semantic level. Relations between colour descriptions are captured precisely. For example, yellow is between red and green in terms of hue, lilac is lighter than purple although they have the same hue. Furthermore, based on the rules of processing adjective modifiers and ranges, we can query in a detailed manner, such as "light blue," which excludes pure blue and dark blue. We can also query on a fuzzy manner, such as "light blue to purple", as required for particular domain purposes.

Furthermore, we have also addressed a common but crucial problem for integration systems: semantic similarities between information from different sources. Two reasonable distance functions are proposed. The distance measuring collaborates with the FaCT-DG DL reasoner to give complete but not redundant results. From our experiments, the simpler distance function (i.e. $d1$) works well enough in a real-world application. By comparing integrated results with their original descriptions, some useful probabilistic conclusions can be inferred, which are especially useful for, e.g., the botanical domain.

Encouraged by the existing results, we plan to extend our work further on ontology-based species identification queries. Firstly, as suggested in Section 6, a future version of our colour reasoner should provide several options so as to allow users to decide on their intended meaning of the "to" keyword. Technically, this requires the use of not only unary but also n-ary datatype expressions as constrains on datatype properties *hasHue*, *hasSaturation* and *hasLightness*. To capture these constraints, we need to use the OWL-E [44,24] ontology language, which is the n-ary extension of OWL-Eu.

Another possible future work is to represent the probabilistic information in the ontology. There are many descriptions with adverbs of quantification, such as "sometimes," "rarely," "often," etc., which also indicate the probability of certain colour information. Because current ontology languages do not support the annotation of classes with probabilities, the probabilistic aspect is ignored in the text processing. This obviously affects the interpretation of integration results. However, there are several attempts to extend DL languages with fuzzy expressions [45,46,47], which, in the future, may be used to enable our logic representation to capture more of the real semantics implied by its original NL descriptions.

Most importantly, from this highly specialised domain, we have learnt a set of more general methodological rules. Key tasks we identified in our study include: (1) modelling the primitive terms (2) based on the semantic model, the effect of modifiers has to be defined and ranges have to be built properly; (3) in order to integrate parallel information, a proper distance measurement is crucial to quantify the similarities among information from multiple sources; (4) depending on the application, more expressive representation and additional reasoning may be necessary to solve real problems. 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. Under this light, many continuous quantities occurring in botany and other descriptive domains, such as leaf shapes, texture, sound, spatial and temporal arrangements, appear to fit fairly straightforwardly into this framework. It is clear that much more development is possible in this very practical area and a holistic system is our future task.

References

1. Clapham, A., Tutin, T., Moore, D.: Flora of the British Isles. Cambridge University Press, Cambridge (1987)
2. Stace, C.: New Flora of the British Isles. Cambridge University Press, Cambridge (1997)
3. Tutin, T.G., Heywood, V.H., Burges, N.A., Valentine, D.H., Moore, D.M. (eds.): Flora Europaea. Cambridge University Press, Cambridge (1993)
4. Gleason, H.: The New Britton and Brown Illustrated Flora of the Northeastern United States and Adjacent Canada. Hafner Publishing Company, New York (1963)
5. Wood, M.M., Lydon, S.J., Tablan, V., Maynard, D., Cunningham, H.: Using parallel texts to improve recall in IE. In: RANLP 2003. Proceedings of Recent Advances in Natural Language Processing, Borovetz, Bulgaria, pp. 505–512 (2003)

6. Wood, M., Lydon, S., Tablan, V., Maynard, D., Cunningham, H.: Populating a database from parallel texts using ontology-based information extraction. In: Meziane, F., Métais, E. (eds.) NLDB 2004. LNCS, vol. 3136, pp. 254–264. Springer, Heidelberg (2004)
7. Wood, M., Wang, S.: Motivation for "ontology" in parallel-text information extraction. In: ECAI-OLP. Proceedings of ECAI-2004 Workshop on Ontology Learning and Population, Poster, Valencia, Spain (2004)
8. Pan, J.Z., Horrocks, I.: OWL-Eu: Adding Customised Datatypes into OWL. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005, LNCS, vol. 3532, Springer, Heidelberg (2005), An extended version is published in the Journal of Web Semantics(to appear)
9. Patel-Schneider, P.F., Hayes, P., Horrocks, I.: OWL Web Ontology Language Semantics and Abstract Syntax. Technical report, W3C, W3C Recommendation (2004)
10. Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Huebner, S.: Ontology-based integration of information - a survey of existing approaches. In: Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, pp. 108–117 (2001)
11. Wang, S., Pan, J.Z.: Ontology-based representation and query colour descriptions from botanical documents. In: Meersman, R., Tari, Z. (eds.) OTM 2005. LNCS, vol. 3761, pp. 1279–1295. Springer, Heidelberg (2005)
12. Lammens, J.M.: A computational model of color perception and color naming. Ph.D. thesis, State University of New York (1994)
13. Berk, T., Brownston, L., Kaufman, A.: A human factors study of color notation systems for computer graphics. Communications of the ACM 25(8), 547–550 (1982)
14. U.S. Department of Commerce, National Bureau of Standards: Color: Universal Language and Dictionary of Names. NBS Special Publication 440. U.S. Government Printing Office, Washington D.C. (1976) (S.D. Catalog No. C13.10:440)
15. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. (eds.): L.A.S.: OWL Web Ontology Language Reference (2004), http://www.w3.org/TR/owl-ref/
16. Pan, J.Z., Horrocks, I.: Extending Datatype Support in Web Ontology Reasoning. In: Meersman, R., Tari, Z., et al. (eds.) ODBASE 2002. LNCS, vol. 2519, pp. 1067–1081. Springer, Heidelberg (2002)
17. Pan, J.Z., Horrocks, I.: Web Ontology Reasoning with Datatype Groups. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, Springer, Heidelberg (2003)
18. W3C Mailing List (starts from 2001), http://lists.w3.org/archives/public/www-rdf-logic/
19. W3C Mailing List (starts from 2004) (2004), http://lists.w3.org/archives/public/public-swbp-wg/
20. Group, J.W.U.P.I.: URIs, URLs, and URNs: Clarifications and Recommendations 1.0., W3C Note (2001), http://www.w3.org/TR/uri-clarification/
21. Fernald, M.: Gray's Manual of Botany. American Book Company, New York (1950)
22. Berk, T., Brownston, L., Kaufman, A.: A new color-naming system for graphics languages. IEEE Computer Graphics and Applications 2(3), 37–44 (1982)
23. Li, L., Horrocks, I.: A Software Framework For Matchmaking Based on Semantic Web Technology. In: WWW 2003. Proc. of the Twelfth International World Wide Web Conference, pp. 331–339. ACM Press, New York (2003)
24. Pan, J.Z.: Description Logics: Reasoning Support for the Semantic Web. PhD thesis, School of Computer Science, The University of Manchester (2004)

25. Gärdenfors, P.: Conceptual Spaces: the geometry of thought. MIT Press, Cambridge (2000)
26. Tversky, A.: Features of similarity. Psychological Review 84(4), 327–352 (1977)
27. Melara, R.: The concept of perceptual similarity: from psychophysics to cognitive psychology. In: Algom, D. (ed.) Psychophysical Approaches to Cognition, pp. 303–388. Elsevier, Amsterdam (1992)
28. Williams, D., Poulovassilis, A.: Combining data integration with natural language technology for the semantic web. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, Springer, Heidelberg (2003)
29. Gaizauskas, R., Wilks, Y.: Information extraction: Beyond document retrieval. Journal of Documentation 54(1), 70–105 (1998)
30. Embley, D., Campbell, D., Liddle, S., Smith, R.: Ontology-based extraction and structuring of information from data-rich unstructured documents. In: Proceedings of International Conference On Information And Knowledge Management, Bethesda, 7, Maryland, USA, (1998)
31. Maedche, A., Neumann, G., Staab, S.: Bootstrapping an ontology-based information extraction system. studies in fuzziness and soft computing. In: Szczepaniak, P., Segovia, J., Kacprzyk, J., Zadeh, L.A. (eds.) Intelligent Exploration of the Web, Springer, Berlin (2002)
32. Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., Shadbolt, N.R.: Automatic ontology-based knowledge extraction from web documents. IEEE Intelligent Systems 18(1), 14–21 (2003)
33. Ferrucci, D., Lally, A.: UIMA: an architectural approach to unstructured information processing in the corporate research environment. Journal of Natural Language Engineering 10(3-4), 327–348 (2004)
34. Goble, C., Stevens, R., Ng, G., Bechhofer, S., Paton, N., Baker, P., Peim, M., Brass, A.: Transparent access to multiple bioinformatics information sources. IBM Systems Journal Special issue on deep computing for the life sciences 40(2), 532–552 (2001)
35. Calvanese, D., Giuseppe, D.G., Lenzerini, M.: Description logics for information integration. In: Kakas, A.C., Sadri, F. (eds.) Computational Logic: Logic Programming and Beyond. LNCS (LNAI), vol. 2408, pp. 41–60. Springer, Heidelberg (2002)
36. Maier, A., Schnurr, H.P., Sure, Y.: Ontology-based information integration in the automotive industry. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 897–912. Springer, Heidelberg (2003)
37. Osgood, C., Suci, G., Tannenbaum, P.: The measurement of meaning. University of Illinois Press, Urbana (1957)
38. Dowty, D.R.: Word Meaning and Montague Grammar. D. Reidel Publishing, Dordrecht (1979)
39. Lakoff, G.: Women, fire, and dangerous things: what categories reveal about the mind. University of Chicago Press, Chicago (1987)
40. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to latent semantic analysis. Discourse Processes 25, 259–284 (1998)
41. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19(1), 17–30 (1989)
42. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: The 32th Annual Meeting of the Association for Computational Linguistics, Las Cruces, Mexico, pp. 133–138 (1994)

43. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: The 14th International Joint Conference on Artificial Intelligence, Montreal, vol. 1, pp. 448–453 (1995)
44. Pan, J.Z.: Reasoning Support for OWL-E (Extended Abstract). In: Basin, D., Rusinowitch, M. (eds.) IJCAR 2004. LNCS (LNAI), vol. 3097, Springer, Heidelberg (2004)
45. Tresp, C., Molitor, R.: A description logic for vague knowledge. In: ECAI 1998. Proceedings of the 13th biennial European Conference on Artificial Intelligence, pp. 361–365. John Wiley and Sons, Chichester (1998)
46. Straccia, U.: Transforming fuzzy description logics into classical description logics. In: Alferes, J.J., Leite, J.A. (eds.) JELIA 2004. LNCS (LNAI), vol. 3229, pp. 385–399. Springer, Heidelberg (2004)
47. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrock, I.: A Fuzzy Description Logic for Multimedia Knowledge Representation. In: Proc. of the International Workshop on Multimedia and the Semantic Web, Crete (2005)