Construction incrémentale d'une ontologie par analyse du texte et de la structure des documents

François Role, Guillaume Rousse

Dans Document numérique 2006/1 Vol. 9 , pages 77 à 91
Éditions JLE

ISSN 1279-5127

Construction incrémentale d'une ontologie par analyse du texte et de la structure des documents

François Role* — Guillaume Rousse**

** Université Paris 5 - René Descartes*
Centre de Recherche en Informatique de Paris 5 (CRIP 5)
45, rue des saints-Pères, F-75006 Paris
francois.role@math-info.univ-paris5.fr

*** INRIA*
Rocquencourt – BP 105
Projet ATOLL, F-78153 Le Chesnay cedex
guillaume.rousse@inria.fr

RÉSUMÉ. *BIOTIM est un projet dont l'objectif est de concevoir des méthodes génériques d'analyse automatique de masses de données regroupant textes et images pour acquérir une sur-couche sémantique commune et, à partir de ce premier résultat, développer des méthodes génériques d'interrogation plurimodale des données ainsi structurées. Dans le cadre de ce projet, nous présentons une expérimentation destinée à améliorer le processus d'acquisition de connaissances grâce à une exploitation simultanée de la structure et du contenu des documents. En particulier, nous montrons comment élaborer une ontologie intermédiaire dans le domaine de la flore tropicale (famille des orchidées) et comment cette ontologie intermédiaire peut contribuer à l'obtention d'une ontologie complète du domaine.*

ABSTRACT. *BIOTIM is a project to design generic methods for the automatic analysis of large amounts of texts and images in order to acquire a common semantic layer and, building upon this initial result, to develop generic methods for a multi-modal examination of the structured data obtained. As part of this project we present an experiment aimed at improving the knowledge acquisition process by exploiting simultaneously both the structure and textual content of documents. In particular, we show how to build an intermediary ontology in the field of exotic flowers (orchidae family) and how this intermediary ontology represents an incremental step in in the building of a complete domain ontology.*

MOTS-CLÉS : *structure logique, ontologie, OWL, traitement automatique du langage naturel, acquisition de connaissances, botanique.*

KEYWORDS: *logical structure, ontologies, OWL, natural language processing, knowledge acquisition, botanics.*

1. Objectifs et motivations

Les progrès des techniques de numérisation et d'édition permettent aujourd'hui de créer des ensembles de ressources de plus en plus volumineux. La capacité à extraire des connaissances à partir de ces « masses de données » est devenue un enjeu très important dans des domaines aussi variés que la recherche scientifique ou la veille économique.

Parmi les documents volumineux pouvant intervenir dans des processus d'acquisition de connaissances, un grand nombre ont une structure logique (manuels techniques, dictionnaires, codes juridiques, etc.) Cette structuration est soit déjà explicitée dans le cas de documents munis d'un balisage XML soit susceptible d'être extraite par des techniques de rétroconversion.

Malgré cette réalité, dans les approches proposées pour la construction d'ontologies à partir de corpus, l'accent est mis sur l'analyse du contenu grâce à des outils d'analyse d'images ou de TALN (Aussenac-Gilles *et al.*, 2000 ; Bourigault *et al.*, 2004) ainsi que sur le nécessaire travail de validation en relation avec des experts du domaine. La contribution que peut apporter l'utilisation de la structure des documents est, elle, rarement explicitée. Assez fréquemment, cette structure est même considérée comme un obstacle à l'analyse du contenu, cette dernière ayant alors comme prérequis des traitements préliminaires destinés à éliminer toute forme de balisage[1].

Une autre source d'informations importante pour l'acquisition automatique de connaissances est constituée par les ressources conceptuelles (ontologies) ou terminologiques (thesaurus, lexiques sémantiques) qui peuvent soit exister dès l'initialisation du processus d'acquisition soit être élaborées à une étape de ce dernier pour être utilisées lors d'étapes ultérieures. Dans ce dernier cas, le processus prend un caractère incrémental puisque des ressources conceptuelles « intermédiaires » contribuent à l'élaboration des ressources conceptuelles qui constituent la cible du processus d'acquisition.

L'élaboration d'une méthodologie d'extraction permettant de faire coopérer efficacement ces sources hétérogènes est un des axes de recherche du projet BIOTIM (BIOTIM, 2003). Dans cet article nous présentons les résultats d'une expérimentation menée dans cette perspective et portant sur l'un des nombreux corpus botaniques actuellement traités dans le cadre du projet BIOTIM.

Après quelques rappels sur le projet BIOTIM et une présentation des caractéristiques des documents ayant servi de support à l'expérimentation (section 2), nous détaillons les étapes du processus d'extraction (section 3).

---

1. On peut bien sûr citer des exceptions comme le système décrit dans (Bourigault *et al.*, 2002) qui s'attache à rechercher des syntagmes nominaux apparaissant dans les intitulés des sections et sous-sections d'un ensemble de codes juridiques, ces syntagmes étant *a priori* pertinents du fait de leur emplacement.

2. Présentation du projet BIOTIM

2.1. *Objectifs du projet*

Labellisé en 2002 dans le cadre de l'ACI « masses de données », le projet BIOTIM a pour objectif de dégager une méthodologie et de développer des outils permettant de construire une ontologie dans le domaine de la flore tropicale. Le projet utilise donc des documents botaniques, mais le but est bien de développer une approche applicable à d'autres types de corpus présentant des caractéristiques similaires en termes de structuration.

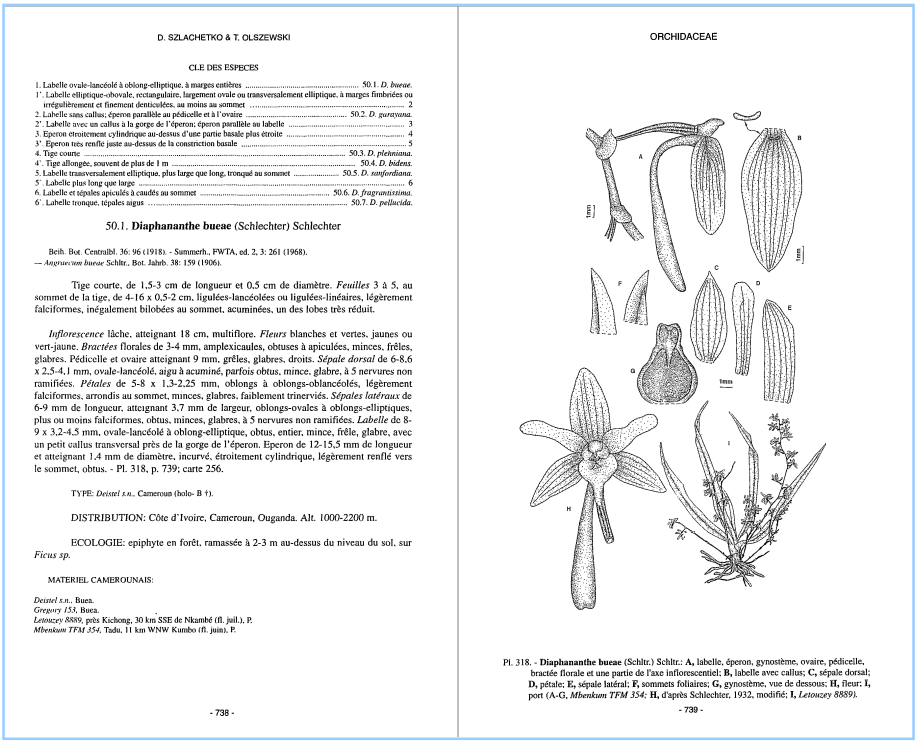
2.2. *Caractéristiques des corpus utilisés*

L'IRD d'Orléans a fourni entre 2003 et 2004 de nombreux corpus numérisés sous forme d'images TIFF (*Flore de la Guyane*, *Flore de la Polynésie*, *Flore du Cameroun*, etc.) aux différents partenaires du projet. Ces derniers exploitent les données fournies à des fins diverses (construction automatique de dictionnaires, analyse automatique des images du corpus, etc.) Cependant l'essentiel des travaux relatifs à l'articulation entre contenu textuel, structure logique et ressources conceptuelles ont été menés au sein de l'équipe ATOLL et ont porté sur la *Flore du Cameroun*, un corpus constitué d'une quarantaine de volumes publiés entre 1963 et 2001, chaque volume représentant environ trois cents pages de texte.

Ce corpus a pour caractéristique d'avoir une structure assez régulière malgré l'étalement dans le temps. Chaque volume est en effet une succession de sections dont chacune décrit un genre ou une espèce. Chaque espèce fait notamment l'objet d'une fiche comportant diverses informations (bibliographie, distribution géographique, écologie, indication des spécimens, etc.) ainsi qu'une partie descriptive qui énumère les caractères distinctifs de l'espèce sous forme de phrases nominales juxtaposées. Cette zone, située au centre de la fiche, contient un grand nombre d'adjectifs qualifiant la forme, la texture, la couleur des organes de l'espèce décrite ainsi que des indications sur le nombre et la dimension de ces organes (figure 1).

2.3. *Premiers essais d'analyse du contenu textuel*

Dès le début du projet, l'idée a été de produire un lexique du domaine de manière à faciliter l'analyse syntaxique devant permettre d'élaborer l'ontologie complète du domaine. La stratégie initialement adoptée a consisté à extraire le texte du corpus de la *Flore du Cameroun*, et à le soumettre à une chaîne d'analyse morpho-syntaxique, le texte ainsi annoté étant alors analysé par des outils d'extraction terminologique.



D. SZLACHETKO & T. OLSZEWSKI

CLE DES ESPECES

1. Labelle ovale-lancéolé à oblong-elliptique, à marges entières ..... 50.1. *D. bueae.*
1'. Labelle elliptique-obovale, rectangulaire, largement ovale ou transversalement elliptique, à marges fimbriées ou irrégulièrement et finement denticulées, au moins au sommet 2
2. Labelle sans callus; éperon parallèle au pédicelle et à l'ovaire ..... 50.2. *D. gurayana.*
2'. Labelle avec un callus à la gorge de l'éperon; éperon parallèle au labelle 3
3. Eperon étroitement cylindrique au-dessus d'une partie basale plus étroite 4
3'. Eperon très renflé juste au-dessus de la constriction basale 5
4. Tige courte ..... 50.3. *D. plehniana.*
4'. Tige allongée, souvent de plus de 1 m ..... 50.4. *D. bidens.*
5. Labelle transversalement elliptique, plus large que long, tronqué au sommet ..... 50.5. *D. tanfordiana.*
5'. Labelle plus long que large 6
6. Labelle et tépales apiculés à caudés au sommet ..... 50.6. *D. fragrantissima.*
6'. Labelle tronqué, tépales aigus ..... 50.7. *D. pellucida.*

50.1. **Diaphananthe bueae** (Schlechter) Schlechter

Beih. Bot. Centralbl. 36: 96 (1918). - Summerh., FWTA, ed. 2, 3: 261 (1968).
— *Angraecum bueae* Schltr., Bot. Jahrb. 38: 159 (1906).

Tige courte, de 1,5-3 cm de longueur et 0,5 cm de diamètre. *Feuilles* 3 à 5, au sommet de la tige, de 4-16 x 0,5-2 cm, ligulées-lancéolées ou ligulées-linéaires, légèrement falciformes, inégalement bilobées au sommet, acuminées, un des lobes très réduit.

Inflorescence lâche, atteignant 18 cm, multiflore. *Fleurs* blanches et vertes, jaunes ou vert-jaune. *Bractées* florales de 3-4 mm, amplexicaules, obtuses à apiculées, minces, frêles, glabres. Pédicelle et ovaire atteignant 9 mm, grêles, glabres, droits. *Sépale dorsal* de 6-8,6 x 2,5-4,1 mm, ovale-lancéolé, aigu à acuminé, parfois obtus, mince, glabre, à 5 nervures non ramifiées. *Pétales* de 5-8 x 1,3-2,25 mm, oblongs à oblongs-oblancéolés, légèrement falciformes, arrondis au sommet, minces, glabres, faiblement trinerviés. *Sépales latéraux* de 6-9 mm de longueur, atteignant 3,7 mm de largeur, oblongs-ovales à oblongs-elliptiques, plus ou moins falciformes, obtus, minces, glabres, à 5 nervures non ramifiées. *Labelle* de 8-9 x 3,2-4,5 mm, ovale-lancéolé à oblong-elliptique, obtus, entier, mince, frêle, glabre, avec un petit callus transversal près de la gorge de l'éperon. Eperon de 12-15,5 mm de longueur et atteignant 1,4 mm de diamètre, incurvé, étroitement cylindrique, légèrement renflé vers le sommet, obtus. - Pl. 318, p. 739; carte 256.

TYPE: *Deistel s.n.*, Cameroun (holo- B †).

DISTRIBUTION: Côte d'Ivoire, Cameroun, Ouganda. Alt. 1000-2200 m.

ECOLOGIE: epiphyte en forêt, ramassée à 2-3 m au-dessus du niveau du sol, sur *Ficus sp.*

MATERIEL CAMEROUNAIS:

Deistel s.n., Buea.
Gregory 153, Buea.
Letouzey 8889, près Kichong, 30 km SSE de Nkambé (fl. juil.), P.
Mbenkum TFM 354, Tadu, 11 km WNW Kumbo (fl. juin), P.

- 738 -

ORCHIDACEAE

Pl. 318. - **Diaphananthe bueae** (Schltr.) Schltr.: **A,** labelle, éperon, gynostème, ovaire, pédicelle, bractée florale et une partie de l'axe inflorescentiel; **B,** labelle avec callus; **C,** sépale dorsal; **D,** pétale; **E,** sépale latéral; **F,** sommets foliaires; **G,** gynostème, vue de dessous; **H,** fleur; **I,** port (A-G, *Mbenkum TFM 354*; **H,** d'après Schlechter, 1932, modifié; **I,** *Letouzey 8889*).

- 739 -

Figure 1. *Description d'une espèce. Les pages des volumes papier fournis par l'IRD ont été numérisées en TIFF, puis le texte a été extrait par un logiciel d'OCR. C'est sur le texte ainsi extrait que portent les travaux d'analyse structurelle et linguistique menés dans le projet BIOTIM*

2.3.1. *Traitement morpho-syntaxique*

La chaîne morpho-syntaxique développée par l'équipe ATOLL a permis d'associer aux corpus du projet BIOTIM, et donc à la *Flore du Cameroun*, des annotations morpho-syntaxiques. Par exemple, la figure 2 fait apparaître les annotations associées par la chaîne morpho-syntaxique au texte « Plantes pourvues d'une tige ». Ces annotations sont représentées en XML conformément à la proposition de normalisation MAF (Clément *et al.*, 2004), ce qui, comme nous le verrons, facilite grandement leur exploitation.

```
<token id="E1F1">Plantes</token>

<wordForm author="lexed" entry="lefff:plante"
form="plantes" tag="cat@noun type@common gender@fem
num@pl" tokens="E1F1"/>

<token id="E1F2">pourvues</token>

<wordForm author="lexed" entry="lefff:pourvoir"
form="pourvues" tag="cat@verb mode@part tense@past
num@pl gender@fem" tokens="E1F2"/>

<token id="E1F3">d'une</token>

<wordForm author="lexed" entry="lefff:de" form="d'"
tag="cat@prep" tokens="E1F3"/>

<wordForm author="lexed" entry="lefff:une" form="une"
tag="cat@det gender@fem num@sing" tokens="E1F3"/>

<token id="E1F4">tige</token>

<wordForm author="lexed" entry="lefff:tige" form="tige"
tag="cat@noun type@common gender@fem num@sing"
tokens="E1F4"/>
```

Figure 2. *Exemple d'annotations morpho-syntaxiques au format MAF*

2.3.2. *Extraction terminologique et essais préliminaires d'acquisition de classes sémantiques*

A l'issue de la chaîne morpho-syntaxique, le texte a été soumis à des outils d'extraction terminologique comme FASTR et ACABIT (Daille, 2003). Dans la phase initiale du projet, l'idée était d'évaluer s'il était possible d'utiliser la terminologie ainsi obtenue pour acquérir des classes sémantiques. Plus précisément, il s'agissait de vérifier si les liens gouverneur-gouverné implicitement présents dans les entrées terminologiques pouvaient servir à effectuer des regroupements sémantiques. Dans l'esprit de la sémantique lexicale, le fait que, par exemple, le terme « teinte » apparaisse comme le gouverneur des termes « jaune », « rouge » et « vert » peut suggérer que ces termes correspondent à un ensemble de couleurs. D'une manière générale, dans ce type d'approche, on essaie d'établir des coefficients de similarité entre les termes en comparant leur contexte. Sur la base des coefficients on calcule une matrice des distances, et il est alors possible d'utiliser un algorithme de classification ascendante hiérarchique pour obtenir un arbre.

Cette technique a donc été expérimentée[2] et des expériences de visualisation sous forme de graphes de termes ont également été menées, mais il est apparu globalement que le résultat des outils d'extraction était bruité, notamment du fait de la non prise en compte de la structuration logique, le contenu de zones particulières comme par exemple la bibliographie, la zone indiquant les références des specimens, l'indication du « type », etc., étant mis sur le même plan que des zones relevant plus directement de la langue naturelle.

La décision a donc été prise de procéder au balisage logique des corpus, de manière à pouvoir mieux cibler les zones à traiter. Il est également apparu qu'il pourrait être intéressant de construire dans un premier temps, non pas une ontologie complète du domaine, mais une ontologie dont la finalité serait de guider les processus d'extraction et d'analyse devant permettre d'obtenir cette ontogie du domaine. Pour tester cette nouvelle approche, il a été décidé de se concentrer sur les trois volumes du corpus de la *Flore du Cameroun* consacrés à la famille des orchidées. Une vingtaine de genres et plus de 360 espèces sont décrites dans ces volumes, la description de la famille des orchidées présentant par ailleurs un intérêt particulier pour les spécialistes du domaine comme le souligne l'équipe éditoriale :

> *« C'est avec joie et soulagement que toute l'équipe responsable de la Flore du Cameroun présente aujourd'hui les Orchidées, une des plus importantes familles de cette série, à la fois par le nombre de ses espèces (360), mais aussi par l'intérêt scientifique, esthétique et économique qu'elle suscite. »* (Szlachetko *et al.*, 2001)

Nous présentons dans les sections suivantes les premiers résultats obtenus en adoptant cette nouvelle approche qui se caractérise donc par une prise en de compte de toutes les informations structurelles disponibles et par l'élaboration d'une ou de plusieurs ontologies intermédiaires.

3. Vers un processus d'extraction incrémental et multisource

L'idée qui guide actuellement les travaux menés au sein de BIOTIM est de construire une ontologie des plantes tropicales en s'appuyant sur la structure du document à deux niveaux :

– d'une part pour obtenir automatiquement, par analyse de cette structure, une hiérarchie de classes reflétant la taxinomie botanique traditionnelle (famille, genre, espèce) ;

2. Le résultat du traitement par ACABIT des 37 volumes de la *Flore du Cameroun* est mis en ligne à l'adresse
http://graves.inria.fr/biotim/resultats/morpho-syntaxe/cameroun/cameroun.acb

– d'autre part pour cibler de manière précise les traitements linguistiques à effectuer pour compléter la hiérarchie de classes par des indications méronymiques portant sur les organes constitutifs des différentes espèces.

Nous allons détailler ces deux étapes (sections 3.2 et 3.3) après avoir donné quelques indications sur le formalisme utilisé pour représenter les ontologies manipulées dans le cadre du projet. (section 3.1).

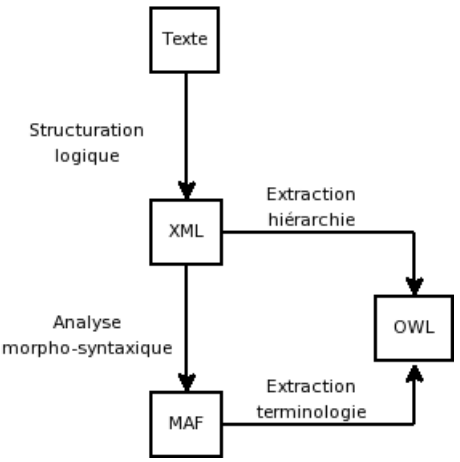


Figure 3. *La structure XML, obtenue par rétroconversion à partir du texte, sert d'une part à construire une hiérarchie de classes OWL reflétant la taxinomie botanique traditionnelle et d'autre part à mieux cibler des traitements linguistiques permettant d'extraire les informations relatives aux organes constitutifs des plantes*

3.1. *Choix du formalisme cible*

La botanique, qui est le domaine traité dans le cadre du projet, se prête naturellement à la représentation de taxinomies. Les plantes sont organisées en genre/espèce et, comme nous le verrons dans la section suivante, la structure logique des documents du corpus reflète directement cette organisation. Il était donc naturel d'utiliser OWL (*Web Ontology Language*) (OWL, 2004), un langage de représentation d'ontologies qui permet de décrire des classes et des instances ainsi que des propriétés que peuvent où doivent posséder des instances. L'utilisation de l'axiome `rdfs:subClassOf` permet de plus d'organiser les classes en une hiérarchie d'héritage. Par ailleurs, OWL existant en trois versions de complexité croissante (*OWL Lite*, *OWL DL*, *OWL Full*) il a été nécessaire d'opérer un choix. *OWL Full* ayant un pouvoir expressif (et une complexité informatique associée)

dépassant les besoins du projet BIOTIM et *OWL Lite* présentant au contraire certaines limitations contraignantes, notamment en termes d'expression des cardinalités, il a donc été décidé de s'en tenir au niveau *OWL DL* pour représenter aussi bien la ou les ontologies intermédiaires que l'ontologie du domaine.

3.2. *Construction d'une hiérarchie de classes OWL à partir de la structure logique*

La première étape de la démarche présentée dans cet article consiste à extraire, grâce à un programme Perl, la structure logique des 37 volumes de la *Flore du Cameroun*[3]. Chaque volume donne ainsi naissance à un document XML, et l'information structurée obtenue est alors utilisée pour amorcer la construction de l'ontologie. À ce stade on dessine en effet déjà les grandes lignes de cette dernière en s'appuyant directement sur la structure macroscopique du document XML (figure 4). Un programme XSLT très simple analyse le document XML et crée une classe OWL pour chaque élément `espèce` rencontré.

```
<genus id="section43" key="51" name="CHAMAEANGIS Schlechter">

<species id="section44" key="51 1" name="Chamaeangis
ichneumonea (Lindley) Schlechter">...</species>

<species id="section45" key="51 2" name="Chamaeangis letouzeyi
Szlachetko">...</species>

<species id="section47" key="51 3" name="Chamaeangis
lanceolata Summerhayes">...</species>

.......

</genus>
```

Figure 4. *Aperçu général de la représentation des relations entre genres et espèces dans un volume de la Flore du Cameroun après conversion en XML*

Le découpage en genres/espèces représenté dans la structure logique de chaque document permet également d'initialiser la hiérarchie de classes de l'ontologie. La figure 5 montre la représentation OWL qui est obtenue pour le genre Diaphananthe. On voit que ce genre descend de la famille des Orchidaceae, et que ses espèces sont représentées par les sous-classes Diaphananthe pellucida, Diaphananthe bueae, etc.

3. Un logiciel d'OCR a permis d'obtenir le texte contenu dans les images de pages similaires à celle qui est représentée sur la figure 1.

```
   <owl:Class rdf:ID="Diaphananthe"/>

   <owl:Class
rdf:about="http://www.flowers.org/ontology#Diaphananthe">

   <rdfs:subClassOf>

   <owl:Class
rdf:about="http://www.flowers.org/ontology#Orchidaceae">

   </owl:Class>

   </rdfs:subClassOf>

   <rdfs:label>Diaphanante</rdfs:label>

   </owl:Class>

   <owl:Class
rdf:about="http://www.flowers.org/ontology#Diaphananthe_pelluc
ida">

   <rdfs:subClassOf>

   <owl:Class
rdf:about="http://www.flowers.org/ontology#Diaphananthe">

   </owl:Class>

   </rdfs:subClassOf>

   </owl:Class>

   <owl:Class
rdf:about="http://www.flowers.org/ontology#Diaphanante_bueae">

   <rdfs:subClassOf>

   <owl:Class
rdf:about="http://www.flowers.org/ontology#Diaphananthe">

   </owl:Class>

   </rdfs:subClassOf>

   </owl:Class>

   .....
```

Figure 5. *Représentation OWL de la hiérarchie de classes*

Dans un second temps, on s'intéresse de manière détaillée aux fiches décrivant les espèces. Comme le montre la figure 6, la structure de ces fiches suit un schéma

assez facile à identifier : bibliographie, zone décrivant les caractéristiques physiques de la fleur, type, distribution géographique et écologie, etc.

Ces informations, fournies directement par la structuration logique, permettent de considérer chaque espèce comme une classe OWL regroupant les individus ayant pour caractéristiques d'avoir une bibliographie associée, une distribution géographique, des spécificités écologiques et un ensemble d'organes. La description précise de ces derniers, qui intéresse en premier lieu les spécialistes du domaine, ne peut cependant être retrouvée qu'en analysant avec des outils linguistiques la zone, située au centre de chaque fiche, qui détaille les caractéristiques physiques de la fleur.

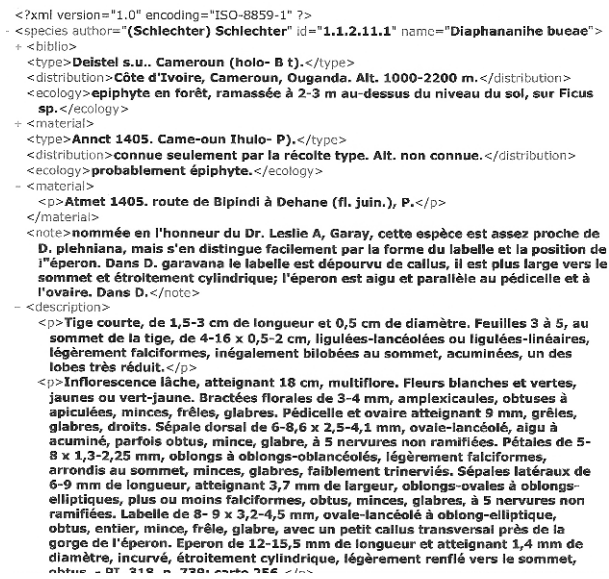


Figure 6. *Représentation XML de la structure logique d'une fiche décrivant une espèce. Les paragraphes situés dans l'élément* `description` *contiennent une description physique détaillée de l'espèce*

3.3. *Coopération entre structure logique et traitements morpho-syntaxiques*

Les techniques de TALN déjà présentées dans la section 2.3 peuvent être mises en œuvre pour contribuer à améliorer la modélisation initiale. En s'appuyant sur la connaissance de la structure, il est déjà possible de cibler les zones les plus appropriées à un traitement linguistique (notamment la zone de texte décrivant les caractéristiques physiques de la fleur, que l'on retrouve au milieu de chaque fiche) et de réserver un traitement à part à des zones ne relevant pas du langage naturel comme la bibliographie par exemple.

Mais on peut faire coopérer encore plus étroitement linguistique et structure en affinant cette dernière. En effet, en nous concentrant sur la zone la plus riche en informations textuelles, à savoir la description des fleurs, l'analyse de cette dernière révèle une structure répétitive qui correspond à une énumération des composantes caractéristiques d'une orchidée, chaque composante faisant l'objet d'une description l'une à la suite de l'autre (ces composantes sont mises en évidence dans le texte de la figure 7).

Tige courte, de 1,5-3 cm de longueur et 0,5 cm de diamètre. **Feuilles** 3 à 5, au sommet de la tige, de 4-16 x 0,5-2 cm, ligulées-lancéolées ou ligulées-linéaires, légèrement falciformes, inégalement bilobées au sommet, acuminées, un des lobes très réduit.

Inflorescence lâche, atteignant 18 cm, multiflore. **Fleurs** blanches et vertes, jaunes ou vert-jaune. **Bractées florales** de 3-4 mm, amplexicaules, obtuses à apiculées, minces, frêles, glabres. Pédicelle et ovaire atteignant 9 mm, grêles, glabres, droits. **Sépale dorsal** de 6-8,6 x 2,5-4,1 mm, ovale-lancéolé, aigu à acuminé, parfois obtus, mince, glabre, à 5 nervures non ramifiées. **Pétales** de 5-8 x 1,3-2,25 mm, oblongs à oblongs-oblancéolés, légèrement falciformes, arrondis au sommet, minces, glabres, faiblement trinerviés. **Sépales latéraux** de 6-9 mm de longueur, atteignant 3,7 mm de largeur, oblongs-ovales à oblongs-elliptiques, plus ou moins falciformes, obtus, minces, glabres, à 5 nervures non ramifiées. **Labelle** de 8-9 x 3,2-4,5 mm, ovale-lancéolé à oblong-elliptique, obtus, entier, mince, frêle, glabre, avec un petit callus transversal près de la gorge de l'éperon. **Eperon** de 12-15,5 mm de longueur et atteignant 1,4 mm de diamètre, incurvé, étroitement cylindrique, légèrement renflé vers le sommet, obtus. - PI. 318, p. 739; carte 256.

Figure 7. *Description des caractéristiques physiques d'une espèce*

Le marquage de cette structure implicite fournit un contexte permettant d'utiliser de manière plus précise les informations morpho-syntaxiques. Il devient ainsi possible d'extraire les qualificatifs utilisés dans la description d'une fleur et de les rattacher aux composantes pour la description desquelles ils peuvent être utilisés. Par exemple, les adjectifs « ligulées-lancéolées », « bilobées », « trilobées » etc. sont adaptés à la description des feuilles d'une fleur alors que « multiflore » ne pourra pas être utilisé dans ce cadre mais sera un bon qualificatif de la composante « inflorescence ».

Nous allons illustrer par un exemple la façon dont ces informations peuvent être obtenues. Pour chaque fiche, on soumet le texte de la zone de description d'une fleur à une une analyse morpho-syntaxique qui produit des annotations similaires à celles qui ont été présentées dans la section 2.3.1. On exploite notamment le fait que l'analyse morpho-syntaxique effectuée par la chaîne de traitement développée par

l'équipe ATOLL est censée déboucher sur une identification non ambiguë des signes de ponctuation dont le point final. De fait, cette identification s'est révélée extrêmement fiable dans le cadre du corpus que nous avons étudié. Ainsi, dans le flot des annotations syntaxiques, un point final est clairement identifié comme dans l'extrait ci-dessous :

```
<token id="E20F3">.</token>

<wordForm author="lexed" entry="lefff:." form="."
tag="cat@ponct" tokens="E20F3"/>
```

On peut alors, par un programme XSLT très simple, structurer finement le document en segmentant la description d'une fleur en zones correspondante à chacune des composantes de cette fleur. On isole par exemple dans une fiche la phrase nominale associée à un sépale dorsal et les annotations morpho-syntaxiques associées à cette composante de la fleur. La figure 8 montre ainsi le début des annotations qui ont été isolées pour la phrase nominale suivante :

« Sépale dorsal ovale-lancéolé, aigu à acuminé, parfois obtus, mince, glabre, à 5 nervures non ramifiées ».

```
<token id="E9F1">Sépale</token>

<wordForm author="lexed" entry="lefff:sépale" form="sépale"
tag="cat@noun type@common gender@masc num@sing"
tokens="E9F1"/>

<token id="E9F2">dorsal</token>

<wordForm author="lexed" entry="lefff:dorsal" form="dorsal"
tag="cat@adj gender@masc num@sing" tokens="E9F2"/>

<token id="E9F19">ovale-lancéolé</token>

<wordForm author="treetagger" entry="talana:_uw" form="ovale-
lancéolé" tag="cat@noun type@common gender@masc num@pl"
tokens="E9F19"/>

<token id="E9F20">,</token>

<wordForm author="lexed" entry="lefff:," form=","
tag="cat@ponct" tokens="E9F20"/>

<token id="E9F21">aigu</token>

<wordForm author="lexed" entry="lefff:aigu" form="aigu"
tag="cat@adj gender@masc num@sing" tokens="E9F21"/>
```

```
<token id="E9F22">à</token>

<wordForm author="lexed" entry="lefff:à" form="à"
tag="cat@prep" tokens="E9F22"/>

<token id="E9F23">acuminé</token>

<wordForm author="lexed" entry="lefff:acuminé" form="acuminé"
tag="cat@adj gender@masc num@sing" tokens="E9F23"/>
```

Figure 8. *Début des* a*nnotations liées à la description d'un sépale dorsal*

Un filtrage linguistique simple basé sur les catégories et les accords permet de retrouver les adjectifs et les participes pouvant qualifier un sépale dorsal :

– « ovale-lancéolé »

– « aigu »

– « acuminé »

– « obtus »

– « mince »

– « glabre »

Par contre le mot « ramifiées » est un participe, mais ses caractéristiques d'accord ne concordent pas avec celles du mot « sépale » et il n'est pas retenu. Au total, un peu plus de 400 qualificatifs applicables aux caractéristiques physiques d'une orchidée tropicale (formes, couleurs, textures) ont pu être intégrés à la description OWL de ce type de fleur, en étant automatiquement assignés aux composantes avec lesquelles ils étaient compatibles.

On obtient ainsi par des traitements linguistiques et structurels très simples (étiquetage morpho-syntaxique, recherche de motifs, détection de la hiérarchie genre/espèce, etc.) une première ontologie « intermédiaire » dans le sens où, même si elle peut déjà présenter un certain intérêt pour les spécialistes de la flore, elle a surtout pour vocation de faciliter des traitements linguistiques plus complexes menés ultérieurement pour élaborer une ontologie complète du domaine. La nécessité d'une telle démarche incrémentale est mise en évidence dans la section suivante.

3.4. *Coopération entre traitements syntaxiques et ressources conceptuelles*

Pour chercher des informations complémentaires dans les descriptions de chaque composante, il est nécessaire d'identifier des relations qui ne se limitent pas aux liens gouverneur-gouverné entre mots proches dans la phrase. Par exemple, dans le texte suivant :

« Sépale dorsal de 6-8,6 x 2,5-4,1 mm, ovale-lancéolé, aigu à acuminé, parfois obtus, mince, glabre, à 5 nervures non ramifiées »

les informations mérologiques (le fait que le sépale possède 5 nervures) ne peuvent être détectées avec l'analyse linguistique de surface présentée dans la section précédente. Une analyse syntaxique plus poussée est nécessaire.

Cependant, il n'est pas en général possible de déboucher sur des analyses non ambiguës en utilisant uniquement des ressources syntaxiques. Par exemple, Dans le cas d'une phrase comme « Feuilles aux nervures épaisses, légèrement acuminées », en l'absence de connaissances sur le domaine, il est impossible de savoir si « acuminées » se rapporte aux feuilles ou aux nervures. L'ontologie « intermédiaires » dont nous avons décrit la construction dans cette section contient précisément ce type d'informations et peut donc contribuer à améliorer la qualité d'une analyse syntaxique plus poussée des fiches botaniques.

Dans les exemples qui précèdent, on voit que l'ontologie doit venir au secours de l'analyse linguistique, mais inversement on peut aussi dire que l'amélioration de l'analyse linguistique contribue à enrichir l'ontologie. Le problème évoqué ici rejoint la question plus générale de la coopération entre analyse syntaxique et ressources conceptuelles, un domaine encore peu exploré comme cela est souligné dans (Sagot *et al*., 2004).

4. Conclusion

Les travaux présentés dans cet article avaient pour but d'évaluer l'intérêt d'une approche consistant à combiner et à faire coopérer étroitement les différentes sources d'information disponibles (contenu, structures documentaires identifiables, ressources conceptuelles – ontologiques ou terminologiques).

Par des traitements très simples (étiquetage morpho-syntaxique, recherche de motifs) il a été possible de produire rapidement une ontologie pouvant déjà servir à piloter des traitements linguistiques complexes (analyse syntaxique et identification de dépendance de complexes). Dans un futur proche, nous serons donc en mesure de construire à partir de cette ontologie intermédiaire une ontologie complète pour la famille des orchidées, ce qui présentera déjà en soi un intérêt pour les spécialistes du domaine. Dans la perspective de fournir à ces spécialistes un environnement de validation motivant et efficace, nous testons déjà en parallèle des outils de diffusion d'ontologies sur le web comme par exemple pOWL ou SESAME. Au total, en nous limitant au cas du corpus étudié, on peut estimer que le souci d'exploiter toutes les informations que pouvaient fournir les données disponibles, notamment les structures textuelles microscopiques et les structures macroscopiques (André *et al.*, 1990), a permis de réduire de façon significative la complexité des programmes à écrire.

Il reste maintenant à vérifier si cette démarche peut être reproduite avec succès pour traiter les autres corpus manipulés dans le cadre du projet BIOTIM[4]. Si c'est le cas, l'exploitation des corpus BIOTIM permettra d'explorer à une très grande échelle la question de la prise en compte d'un point de vue linguistique des ressources conceptuelles et structurelles.

5. Bibliographie

ATOLL, Projet ATOLL, Inria-Rocquencourt, http://atoll.inria.fr/

André J., Quint V., « Structures et modèles de documents », *in Le document électronique : cours INRIA*, Châtelaillon, 11-15 juin 1990, p. 3-57.

Aussenac-Gilles N., Biebow B., Szulman S., « Corpus analysis for conceptual modelling », *Workshop on Ontologies and Texts, Knowledge Engineering and Knowledge Management: Methods, Models and Tools, 12th International Conference, EKAW'2000*, Juan-les-pins, France, octobre 2000.

BIOTIM, Projet ACI BIOTIM, http://www-rocq.inria.fr/imedia/biotim/

Bourigault D., Aussenac-Gilles N., Charlet J., « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur », *RIA*, vol. 18, n° 1, 2004, p. 87-110.

Bourigault D., Lame G., « Analyse distributionnelle et structuration de terminologie. Application à la construction d'un ontologie documentaire du droit », *TAL*, 43, 2002, p. 129-150.

Clément L., Villemonte de la Clergerie E., Terminology and other language resources – morpho-syntactic annotation framework (MAF), ISO TC37 SC4 WG2 Working Draft, 2004.

Daille B., « Terminology mining », *Information Extraction in the Web Era*, *Lectures Notes in Artificial Intelligence*, 2003, p. 29-44.

OWL, Web Ontology Language (OWL), http://www.w3.org/2004/OWL/

Sagot B., El Ghali A., « Coupling grammar and knowledge base: Range Concatenation Grammars and Description Logics », *in Lecture Notes in Artificial Intelligence 3206, Proceedings of TSD'04*, Brno, 2004, p. 195-202.

Szlachetko D., Olszewski T., *Flore du Cameroun,* vol. 34, ministère de la Recherche scientifique et technique du Cameroun, Yaoundé, 2001.

4. On peut sans trop s'avancer, supposer que c'est au moins le cas pour d'autres familles de fleurs, la description de ces dernières étant stéréotypée d'une famille à l'autre.