International Journal on Artificial Intelligence Tools
© World Scientific Publishing Company

CONCEPTUAL GRAPHS AND SYSTEMS BIOLOGY

M. VILARES FERRO and M. FERNÁNDEZ GAVILANES and A. BLANCO GONZÁLEZ

Dept. of Computer Science, University of Vigo, Campus As Lagoas s/n
32004 Ourense, Spain
{vilares,mfgavilanes,adbgonzalez@}uvigo.es

C. GÓMEZ RODRÍGUEZ

Dept. of Computer Science, University of A Coruña, Campus de Elviña s/n
15071 A Coruña, Spain
cgomezr@udc.es

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

A knowledge discovery and representation frame to mine contents in systems biology is described. It applies natural language processing to integrate linguistic and domain knowledge in a mathematical model for information management, formalizing the notion of semantic similarity in different degrees. The goal is to provide computational tools to identify, extract and relate not only data but also scientific notions, even if the information available to start the process is not complete.

The interpretation basis is the conceptual graph, a formalism for semantic representation that allows us to express meaning in a form that is logically precise, humanly readable, and computationally tractable. Our work exploits the automatic generation of these structures from raw texts through graphical and natural language interaction, providing a solid foundation for the treatment of document incompleteness and query vagueness. We avoid recourse to classic ontologies serving as meta languages for the annotation task that frequently prevent the effective reuse of knowledge, unnecessarily overloading the accessing task for inexpert users, which significantly distances us from previous approaches.

Keywords: Conceptual graphs; knowledge acquisition; knowledge mining; natural language processing; systems biology.

2016. DOI 10.1142/S0218213015500293

1. Introduction

The globalization of information access and the generation of vast amounts of knowledge in the biological domain have provided the ground for a new research paradigm called *systems biology* (SB) [35]. In contrast to the classic hypothesis-oriented approach, where the researcher attempts to refute hypotheses from a set of baseline assumptions, this new one focuses on the integration in a single model of complex interactions between large amounts of heterogeneous data to yield specific targets for further research in the sense of the data-oriented experimental paradigm described in [53]. Such an approach is especially interesting when dealing with very labor-intensive and time-consuming tasks, involving the analysis, crosscheck and synthesis of large amounts of data, as is the case of phenotype identification or seeking an understanding of the evolution, working and interactions of biological organisms. As it is commonly accepted that only 10-20% of all species have been discovered [55], it can be expected that the current data overload in biodiversity will continue to grow, making it impossible for researchers to read all the literature that would be of interest to solve any of these specific problems. This involves a critical need for information discovery tools, whose economical impact must be taken into consideration [9], posing a number of problems that need to be overcome.

A first question to address is how to represent the huge volume of knowledge we want to extract and manage. As most biological data are currently available in digital format [42] or are currently being digitized, research to provide semantic-based access to information naturally leads us to the notion of ontology, intended as a framework for the domain knowledge of an intelligent system [12]. This justifies the popularity of this kind of structures during recent years [27] in a variety of biological domains [15,30,56], although shortcomings in both ontology content and logic design often prevent the creation of operational frameworks [18]. In particular, as one major source of data is scientific publications in the form of raw text using natural language, creating such curated structures often relies on qualified experts that not only need to provide access to relevant information but also to inter-relate it while ensuring collaborative sharing and maintainability. All this once again incurs high costs in terms of both time and staff.

Another problem raised concerns the access to information itself. Although the web has revolutionized this technology over the last decade by providing a distributed architecture linking frameworks and servers [3], most authoritative databases continue to be accessed through traditional entry points [55], whilst the meaning of documents is left unanalyzed and can only be explored using full-text techniques [26]. This means some limitations for transparent, precise and easy retrieval by inexpert users across heterogeneous data sources, as well as for flexible and rapid adaptation to changes in the domain [37]. To solve these without incurring an information overload, we need to enable strategies for both *knowledge mining* (KM) and *discovery*

(KD) on unstructured documents expressed in natural language [1].

Finally, the interactive nature of the actual knowledge management task must be factored. On the one hand, biologists need to relate information from different and heterogeneous sources [52], as is the case of diagrammatic biological models or phenotypical, geographical or experimental data. On the other hand, the researcher should be able to estimate the accuracy, efficiency and reliability of the system [41]. However, a black box device fails to capture the user's true context and problem solving cannot go beyond its own knowledge resources, even if this is possible from a logical point of view. As a consequence, the behavior of such a system turns out to be transparent and uniform across all users [53].

On the basis of the remarks made in the foregoing paragraphs, we propose a frame for KD and KM taking the notion of *conceptual graph* (CG) [50] as representation formalism. This makes it possible to identify semantic structures with formulae, and introduce deduction as reasoning model through a graph morphism called *projection* [11], which proves to be both sound and complete with regard to deduction in *first-order-logic* (FOL), thereby providing a formal frame for flexible information access [23]. With regard to user interaction, we are interested in dealing directly with raw text, which allows for the immediate treatment of most information resources at low maintenance costs regardless of their nature and complexity. In particular, we propose an interface design based on *natural language processing* (NLP) that benefits from the semantic analysis capabilities of our KM architecture through a deep dependency parse [17]. This latter permits both a high degree of grammatical abstraction and the integration of constituency information modeled by a *tree-adjoining grammar* (TAG) [33], which facilitates the rapid extraction of grammatical relations, while we gain in robustness and the complexity remains polynomial. Finally, the use of this parse as a starting point for the generation of the CGs that support our semantic representation is the foundation we need to implement an alternative graphical interface.

2. The state-of-the-art

The emergence of SB as research paradigm has revolutionized the biological domain, arousing the interest for specific techniques in knowledge representation, data integration, querying and hypothesis generation [3]. Intended as a computer-understandable model for knowledge representation, ontologies have proved to be a powerful tool for conceptual specification, holding great potential for data sharing and reusing [4], and also providing the necessary support for formal reasoning [14]. In this sense, the biological community has devoted great efforts in recent years to their creation and maintenance, mainly under the umbrella of the *Open Biological Ontologies Foundry* (OBO) [49], a collaborative experiment that has resulted in a set of principles for their development in the domain. The initiative includes more than a hundred ontologies on the various branches of biology such as anatomy, behavior, environmental conditions, experimentation, phenotype, process, sequence or taxonomy.

Such a variety of data sources impacts both the reasoning task on computable definitions and its maintainance, due to the compositional nature of the information, thus favoring links between concepts in different ontologies [40]. This, in turn, can lead to redundancy phenomena in both concepts and relationships, with a consequent increase in complexity. Data interoperability and integration then arises as a major research topic [7], with different lines of work open, many of them already available within the OBO library.

The most simple way of approaching the problem described is the use of servers publishing ontology repositories [19] or ontology frameworks [10], the latter understood as collections of functions to access ontologies that have probably been stored in distinct formats. It is also possible to integrate both approaches [16], although they are sensitive to update processes, complicating the maintenance task for accessing tools as well as contents. In order to reduce this impact, most authors are focusing their efforts on two types of strategies. The former involves the development of terminologies serving as meta-language for the annotation tasks, providing a shared vocabulary [36] or an automated annotation system [14], whilst the second one looks for upper domain ontologies supporting knowledge sharing and management [5,52]. Unfortunately, none of these alternatives show the flexibility that could have been expected when the data handled are inherently heterogeneous, mostly written in natural language, provided by different researchers possibly using distinct vocabularies and methodologies, pursuing particular goals and under varying spatio-temporal frames [16].

The querying task attempts to adapt to this rigidity by incorporating natural language interfaces whose interpretation is guided by a domain ontology, providing a mapping between linguistic structures and domain conceptual relations [39], often through a dependency parse [6,37,53]. This allows us to enter structural matching to support similarity measures between concepts [20] and clears the way for hypothesis generation, automating the knowledge acquisition task from raw texts. However, addressing the query problem effectively lies first in locating it in a decidable framework that, in turn, also impacts on the rest of techniques relating the computational treatment of SB. In this sense, CGs [50] provide a tool to make commonalities explicit and to derive knowledge, inferring sub and super-concept relationships. The application of the model to KM has become of increasing interest in the last decade [8,29,58], although practical proposals have been limited to restricted domains, looking for both grammatical simplicity and a well-known vocabulary in texts: patent claims [43], financial statements [34] or computer science [38,47]. The initial cause of these limitations is the difficulty to automatically generate CGs from raw texts, entailing a manual processing of the documents. At this point, attempts to solve this question have been rare and incomplete [28]. To the best of our knowledge, no practical proposals in the biological domain are available and only some preliminary work has suggested their applicability in this case [24], probably due to its inherent linguistic complexity.

With this aim in mind, we describe a practical framework that makes it possible to complete the automation of the KM process, saving time and resources. Introduced by taking a botanic *corpus* as documentary collection, the technique illustrates the applicability of CG-based approaches to also deal with the biodiversity domain. The proposal includes both a natural language interface and a graphical one, the latter enabling the direct management of CGs in querying tasks, which allows the user to fully exploit this FOL-based deduction model.

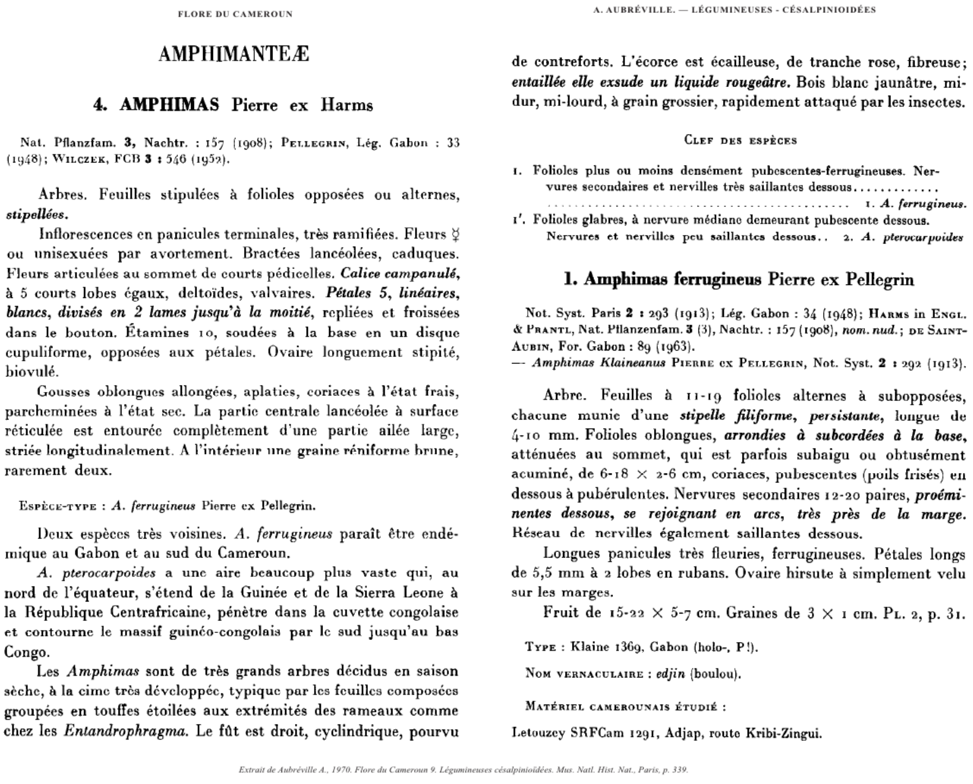
3. The running corpus

We introduce our proposal from a botanic *corpus* describing West African flora: the *"Flore du Cameroun"*, published between 1963 and 2001, drawn by different research groups and supplied by the *French Institute of Research for Cooperative Development.* It consists of about forty volumes in French, each one running to about three hundred pages. The text is organized taxonomically, introducing genera (resp. species) in separate chapters (resp. sections), and the descriptions include concepts that are related both taxonomically and non-taxonomically. In the first case, concepts are organised into sub and super tree structures, involving the logic relationships which seem to be the most frequent in biological ontologies [13], namely the generic, partitive and instance ones. The former is often formalized by "is-a", as in the description referred to the genus Afzelia, where the sentence *"l'Afzelia bipidensis existe depuis le sud du Nigeria jusqu'a l'Angola"* ("the Afzelia bipidensis grows from Nigeria to Angola") corresponds to an annotation of the type "Afzelia bipidensis is-a Afzelia". The most common form for the second is "part-of", as in *"feuilles à nervures"* ("leaves with veins"), which associates an annotation of the type "vein part-of leaf". For its part, instances describe individuals, as in *"de couleur rose"* ("of pink color"), that we can formalize by "pink instance-of color".

Non-taxonomic relations include equivalence and associative links. The first relate to concepts that can be represented by more than one entry, which is not unusual in a domain where different names can be assigned to a same organism at different times, either by mistake or due to the existence of vernacular names in use [52]. An example could be the expression *"noms vernaculaires: radis"* ("vernacular names: radish") included in the section devoted to the species *Raphanus sativus.* The associative case involves thematic links between terms that are neither hierarchical nor equivalent, but are nevertheless semantically or contextually related to one another. So, concepts can be siblings within the same branch of the hierarchy or exist in separate ones, baptized as related terms and cross-references respectively. These relationships are reversible, imply inheritance and often arise from experience, providing context and meaning to make the reasoning work easier, for example in dealing with *information retrieval* (IR) tasks through query expansion [44]. Focusing on biology, they commonly fall into four broad kinds [51], which we try to contextualize within a reference framework in this domain of knowledge, the *Plant Ontology* (PO) [30] database:

- Nominative relationships describing the names of concepts, as is the case in *”lobes latéraux appelés rostellophores”* ("lateral lobes called rostellophores”), captured by "lateral lobe hasName rostellophore”. Currently, PO does not provide this kind of facility.
- Locative relationships, referring the location of one concept with respect to another as in *”fente plus profonde à côté des étamines”* ("deeper slot next stamens”), which can be captured using annotations of the type "deeper slot adjacent-to stamens”. Another example can be *”sores localisés dans les sinus”* ("sori located in the sinuses”), corresponding to an annotation of the type "sorus located-in sinus”.
- Relationships representing the functions (resp. processes) a concept has (resp. is involved in), and other properties of the concept. It is the case of *”les deux étamines stériles se projettent devant le stigmate, peut-être, jouent un rôle de déclencheur dans la pollinisation”* (resp. *”racines aérifères spongieuses ayant un rôle de pneumatophores”*) ("two sterile stamens projected in front of the stigma, perhaps playing a triggering role in pollination”) (resp. "air springing spongy roots with a role of pneumatophores”), which we can formalize by "two sterile stamens in front of the stigma participate-in pollination” (resp. "pneumatophore develops-from air springing spongy roots”). Although PO also includes a relation "derives-by-manipulation-from”, whose reverse is "has-participant”, this is reserved to relate *in vitro* plant structures, none of which are represented in our *corpus.*
- Other types of links, such as temporal ones in *”les fruits mûrissent dès janvier”* ("the fruits ripen in January”) that we can characterized by "ripe fruits co-occurs-with January”. To date, they are not considered by PO.

When all these relations between concepts rely on a concrete application domain, eventually enriched with other links, they form a *domain ontology.* This provides a map between knowledge and linguistic data, which justifies its use by NLP tools in different contexts, such as IR systems, to improve expressiveness and accuracy [37]. Our aim is to serve the same purpose, identifying and accessing the same relations and concepts, but taking the information directly from the unstructured text. However, in order to illustrate this kind of process, the *corpus* should not only include the semantic information needed to generate those structures, as seen above, but it should also do so involving resources of sufficient size and complexity in both grammatical and lexical aspects. In this regard, each chapter is organized in sections with a title, a narrative description and a dichotomy. A section can also incorporates subsections following the same structure. As an overall rule, the title includes in its first line the names of the authors and the taxon family and subfamily we are dealing with. A second line refers the botanical genus to which the section is devoted, as well as the author who made the discovery. Descriptions relate to morphological aspects such as color, texture, size or form. This implies the presence of nominal sentences, adjectives and also adverbs to express frequency and intensity, and named entities to denote dimensions. A set of dichotomical keys is included when the range presented has other inferior ones. An example with a fragment of one of these sections is shown in Fig. 1, related to the *Amphimas* genus[a]. Finally, the grammatical structure is added to enable us to propagate the relationships through linguistic constructions, as with enumerations on expressions pointing out instances for the color or the form. Such is the case of *”écorce lenticellée de couleur grisâtre, brunâtre ou pourprée”* ("lenticellate bark of greyish, brownish or purple color”) and *”limbe .... de forme très variées, entier ou profondément digité ou palmatisêqué, ...”* ("limb ... of very varied shape, entire or deeply fingered or palmatisected, ...”), respectively.



FLORE DU CAMEROUN

AMPHIMANTEÆ

4. AMPHIMAS Pierre ex Harms

Nat. Pflanzfam. **3,** Nachtr. : 157 (1908); PELLEGRIN, Lég. Gabon : 33 (1948); WILCZEK, FCB **3 :** 546 (1952).

Arbres. Feuilles stipulées à folioles opposées ou alternes, ***stipellées.***

Inflorescences en panicules terminales, très ramifiées. Fleurs ☿ ou unisexuées par avortement. Bractées lancéolées, caduques. Fleurs articulées au sommet de courts pédicelles. ***Calice campanulé,*** à 5 courts lobes égaux, deltoïdes, valvaires. ***Pétales 5, linéaires, blancs, divisés en 2 lames jusqu'à la moitié,*** repliées et froissées dans le bouton. Étamines 10, soudées à la base en un disque cupuliforme, opposées aux pétales. Ovaire longuement stipité, biovulé.

Gousses oblongues allongées, aplaties, coriaces à l'état frais, parcheminées à l'état sec. La partie centrale lancéolée à surface réticulée est entourée complètement d'une partie ailée large, striée longitudinalement. A l'intérieur une graine réniforme brune, rarement deux.

ESPÈCE-TYPE : *A. ferrugineus* Pierre ex Pellegrin.

Deux espèces très voisines. *A. ferrugineus* paraît être endémique au Gabon et au sud du Cameroun.

A. pterocarpoides a une aire beaucoup plus vaste qui, au nord de l'équateur, s'étend de la Guinée et de la Sierra Leone à la République Centrafricaine, pénètre dans la cuvette congolaise et contourne le massif guinéo-congolais par le sud jusqu'au bas Congo.

Les *Amphimas* sont de très grands arbres décidus en saison sèche, à la cime très développée, typique par les feuilles composées groupées en touffes étoilées aux extrémités des rameaux comme chez les *Entandrophragma.* Le fût est droit, cylindrique, pourvu

A. AUBRÉVILLE. — LÉGUMINEUSES - CÉSALPINIOIDÉES

de contreforts. L'écorce est écailleuse, de tranche rose, fibreuse; ***entaillée elle exsude un liquide rougeâtre.*** Bois blanc jaunâtre, mi-dur, mi-lourd, à grain grossier, rapidement attaqué par les insectes.

CLEF DES ESPÈCES

1. Folioles plus ou moins densément pubescentes-ferrugineuses. Nervures secondaires et nervilles très saillantes dessous.............
   .... 1. *A. ferrugineus.*

1′. Folioles glabres, à nervure médiane demeurant pubescente dessous.
   Nervures et nervilles peu saillantes dessous.. 2. *A. pterocarpoides*

1. Amphimas ferrugineus Pierre ex Pellegrin

Not. Syst. Paris **2 :** 293 (1913); Lég. Gabon : 34 (1948); HARMS in ENGL. & PRANTL, Nat. Pflanzenfam. **3** (3), Nachtr. : 157 (1908), *nom. nud.*; DE SAINT-AUBIN, For. Gabon : 89 (1963).
— *Amphimas Klaineanus* PIERRE ex PELLEGRIN, Not. Syst. **2 :** 292 (1913).

Arbre. Feuilles à 11-19 folioles alternes à subopposées, chacune munie d'une ***stipelle filiforme, persistante,*** longue de 4-10 mm. Folioles oblongues, ***arrondies à subcordées à la base,*** atténuées au sommet, qui est parfois subaigu ou obtusément acuminé, de 6-18 × 2-6 cm, coriaces, pubescentes (poils frisés) en dessous à pubérulentes. Nervures secondaires 12-20 paires, ***proéminentes dessous, se rejoignant en arcs, très près de la marge.*** Réseau de nervilles également saillantes dessous.

Longues panicules très fleuries, ferrugineuses. Pétales longs de 5,5 mm à 2 lobes en rubans. Ovaire hirsute à simplement velu sur les marges.

Fruit de 15-22 × 5-7 cm. Graines de 3 × 1 cm. PL. 2, p. 31.

TYPE : Klaine 1369, Gabon (holo-, P!).

NOM VERNACULAIRE : *edjin* (boulou).

MATÉRIEL CAMEROUNAIS ÉTUDIÉ :

Letouzey SRFCam 1291, Adjap, route Kribi-Zingui.

Extrait de Aubréville A., 1970. Flore du Cameroun 9. Légumineuses césalpinioïdées. Mus. Natl. Hist. Nat., Paris, p. 339.

Fig. 1. The description of Amphimas genus

The vocabulary is shared by most texts on this matter and, due to the diversity of the constructions present in the *corpus* and the different ways in which they are

[a] it can be recovered from `http://phyto-afri.ird.fr/Flore_cameroun/amphimas.pdf`

expressed, it also seems to be a suitable testing platform to deal with ambiguity and grammatical completeness. We denote this *corpus* by $\mathscr{B}$, whose main data set features are a size of 33.9 Gb with 2,719 documents that include a total of 863,297 terms. When it comes to document length, the minimum (resp. maximum) size is 15 (resp. 58,297) words, the average length being 2,079.46.

4. Conceptual graphs and searchable bases

That follows is a survey of the key concepts in CG theory and the formalization of the question answering problem, both necessaries to understand our proposal. Most of the them are taken from [11,23].

Definition 1. A *support* is a triple $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$ of finite sets pairwise disjoint, such that $\mathcal{T_C}$ (resp. $\mathcal{T_R}$) is a partially ordered set of *concept* (resp. *relation*) *types.* These orders are interpreted as specialization relationships. So, $t \leq r$ is read as r *is a generalization of* t or, also, as t *is subsumed by* r. Types in $\mathcal{T_C}$ posses a greater element, $\top$, called *universal type.* Types in $\mathcal{T_R}$ may be of any arity greater or equal to 1, and only those with same arity are comparable. The countable set $\mathcal{I}$ is a collection of *individual markers* with a *generic marker* $* \notin \mathcal{I}$. The set $\mathcal{I} \cup \{*\}$ is partially ordered and its elements pairwise non-comparable, being $*$ the greatest one.

A *support* compiles the main concepts, relations and vocabulary that exist in the world we are trying to describe. We can identify the set of markers with a dictionary representing the lexical forms we are working in, while concepts refer to their semantic categories and relations the relationships between them.

Definition 2. A *basic conceptual graph (*BG*)* defined over a support $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$ is a 4-tuple $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{E}, \mathcal{L})$, where $(\mathcal{C} \cup \mathcal{R}, \mathcal{E})$ is a bipartite multigraph with $\mathcal{C}$ and $\mathcal{R}$ disjoint sets of *concept* and *relation nodes*, respectively. $\mathcal{E}$ is the multiset of *edges* and $\mathcal{L}$ is a labeling function for nodes and edges. A node $c \in \mathcal{C}$ is labeled by a pair $[\mathit{type(c)}, \mathit{marker(c)}] \in \mathcal{T_C} \times (\mathcal{I} \cup \{*\})$. A node $r \in \mathcal{R}$ is labeled by $\mathit{type(r)} \in \mathcal{T_R}$ and the degree of r, i.e., the number of edges incident to, must be equal to the arity of *type(r)*. An edge in $\mathcal{E}$, labeled by $i \in \mathbb{N}$, connecting nodes $r \in \mathcal{R}$ and $c \in \mathcal{C}$, is denoted by (r, i, c). The edges $(r, 1, c_1), \ldots, (r, k, c_k)$ incident to $r \in \mathcal{R}$ are totally ordered and labeled from 1 to the degree k of r. We then shortly denote $r = \mathit{type(r)}(c_1, \ldots, c_k)$.

Essentially, a BG is a CG without negation, which simplifies our description. It represents the stencil to be filled in with the concept/relations taken from the support, providing an ontology of the domain.

Definition 3. Let $\mathcal{G}_1 = (\mathcal{C}_1, \mathcal{R}_1, \mathcal{E}_1, \mathcal{L}_1)$ and $\mathcal{G}_2 = (\mathcal{C}_2, \mathcal{R}_2, \mathcal{E}_2, \mathcal{L}_2)$ be BGs defined on a support $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$, then a *projection* from $\mathcal{G}_1$ to $\mathcal{G}_2$ is a mapping π from $\mathcal{C}_1$ to $\mathcal{C}_2$, and from $\mathcal{R}_1$ to $\mathcal{R}_2$ verifying:

$$(r, i, c) \in \mathcal{E}_1 \Rightarrow (\pi(r), i, \pi(c)) \in \mathcal{E}_2 \quad \mathit{and} \quad x \in \mathcal{C}_1 \cup \mathcal{R}_1 \Rightarrow \mathcal{L}_2(\pi(x)) \leq \mathcal{L}_1(x) \qquad (1)$$

where, if $x \in \mathcal{C}_1$, $\leq$ refers to the cartesian product of the order on $\mathcal{T_C}$ and on $\mathcal{I} \cup \{*\}$[b]. If $x \in \mathcal{R}_1$, then $\leq$ refers to the order on $\mathcal{T_R}$. We call $\mathcal{G}_1$ the *source* and $\mathcal{G}_2$ the *target*, and we say that $\mathcal{G}_1$ *subsumes* $\mathcal{G}_2$ or that $\mathcal{G}_1$ is *more general than* $\mathcal{G}_2$, using the notation $\mathcal{G}_1 \succeq \mathcal{G}_2$. The set of projections from $\mathcal{G}_1$ to $\mathcal{G}_2$ is denoted by $proj(\mathcal{G}_1, \mathcal{G}_2)$.

A projection is a graph homomorphism that can specialize the labels of concept and relation nodes. So, the existence of a projection from a BG $\mathcal{Q}$ to another one $\mathcal{D}$ means that the knowledge represented by $\mathcal{Q}$ is contained in the one represented by $\mathcal{D}$, defining a comprehensive indexation protocol for IR purposes.

Theorem 1. Let $\mathcal{G}_1 = (\mathcal{C}_1, \mathcal{R}_1, \mathcal{E}_1, \mathcal{L}_1)$ and $\mathcal{G}_2 = (\mathcal{C}_2, \mathcal{R}_2, \mathcal{E}_2, \mathcal{L}_2)$ be BGs defined on a support $\mathcal{S}$, then $\mathcal{G}_1 \succeq \mathcal{G}_2$ iff $\exists\, \pi$, a projection from $\mathcal{G}_1$ to $\mathcal{G}_2$.

Proof 1. Trivial from Definition 3.

This introduces a basic querying mechanism, although to locate answers that do not exactly correspond to projections, which often happens in the real world, we must relax the structural constraints.

Definition 4. Let $\mathcal{D}$, $\mathcal{D}'$ and $\mathcal{Q}$ be BGs defined on a support $\mathcal{S}$, and ς a mapping from the set of BGs defined on $\mathcal{S}$ onto itself, such that $\varsigma(\mathcal{D}) = \mathcal{D}'$. If $\pi \in proj(\mathcal{Q}, \mathcal{D}')$, then (π, ς) is a *projection from $\mathcal{Q}$ to $\mathcal{D}$ modulo ς*.

The idea is to supply a set of transformations in order to determine the degree of relevance of a document to a query, when some kind of relation links them. The greater the structural impact on the document, the less relevance will be with regard to the query.

Definition 5. Let $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{E}, \mathcal{L})$ be a BG defined on a support $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$, a *substitution on $\mathcal{G}$* is a pair $(t, t') \in (\mathcal{C} \times (\mathcal{T_C} \times (\mathcal{I} \cup \{*\}))) \cup (\mathcal{R} \times \mathcal{T_R})$. When we assert that the concept (resp. relation) term t can replace the term t', we say that (t, t') are *compatible terms*.

Definition 6. Let $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{E}, \mathcal{L})$ be a BG defined on a support $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$, the result of the *join of* $c, c' \in \mathcal{T_C}$, such that $\mathcal{L}(c) = \mathcal{L}(c')$, is the BG obtained from $\mathcal{G}$ by identification of c and c'.

As a join can substantially change the structure of a BG, this transformation is usually considered more distancing than substitutions.

Definition 7. Let $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{E}, \mathcal{L})$ be a BG defined on a support $\mathcal{S} = (\mathcal{T_C}, \mathcal{T_R}, \mathcal{I})$, the result of *adding a node* $n \in \mathcal{C} \cup \mathcal{R}$, such that $\mathcal{L}(n) = v$, is the new BG $\mathcal{G} + \mathcal{N}$, where $\mathcal{N}$ is the graph reduced to n. If $n \in \mathcal{R}$, neighbors must be specified.

[b] i.e., $(type(\pi(x)), marker(\pi(x))) \leq (type(x), marker(x))$ iff $type(\pi(x)) \leq type(x)$, and $marker(\pi(x)) \leq marker(x)$.

Given that an addition modifies the BG, but also introduces an external element, it is taken to be more complex than a join. According to the combination of these transformations, we consider four kinds of answer to a given query.

Definition 8. Let $\mathcal{D}$ and $\mathcal{Q}$ be BGs defined on a support $\mathcal{S}$, $\mathcal{D}$ is an *exact answer to* $\mathcal{Q}$ iff *proj*$(\mathcal{Q}, \mathcal{D}) \neq \emptyset$.

Since the absence of an exact answer is usual, either as a result of the lack of information in the documentary database (*documentary incompleteness*) or in the query itself (*query vagueness*), we need to capture the notion of non-exact one.

Definition 9. Let $\mathcal{D}$ and $\mathcal{Q}$ be BGs defined on a support $\mathcal{S}$. Then $\mathcal{D}$ is an *approximate answer to* $\mathcal{Q}$ iff there exists a sequence of substitutions ς, such that *proj*$(\mathcal{Q}, \varsigma(\mathcal{D})) \neq \emptyset$.

As exact answers are a rare case of approximate ones, we use this last term to refer both categories. In order to further increase the degree of flexibility associated to querying, we can also include joins as admissible transformations.

Definition 10. Let $\mathcal{D}$ be a BG defined on a support $\mathcal{S}$. We say that a *sequence* ς *of substitutions and joins is acceptable* iff ς does not contain too many joins relative to the number of nodes in the BG $\mathcal{Q}$. The ratio number (μ_j) of joins is chosen by the user.

Definition 11. Let $\mathcal{D}$ and $\mathcal{Q}$ be BGs defined on a support $\mathcal{S}$. We say that $\mathcal{D}$ is a *plausible answer* to $\mathcal{Q}$ iff there is an acceptable sequence ς of substitutions and joins, such that *proj*$(\mathcal{Q}, \varsigma(\mathcal{D})) \neq \emptyset$.

We finally include node adds in order to complete the range of possible transformations.

Definition 12. Let $\mathcal{D}$ be a BG defined on a support $\mathcal{S}$. We say that a *sequence* ς *of substitutions, joins and node adds is acceptable* iff ς is acceptable for the joins and there are not too many node adjunctions relative to the number of nodes in the BG $\mathcal{Q}$. The ratio number (μ_a) of node added is chosen by the user.

Definition 13. Let $\mathcal{D}$ and $\mathcal{Q}$ be BGs defined on a support $\mathcal{S}$. We say that $\mathcal{D}$ is a *partial answer* to $\mathcal{Q}$ iff there is an acceptable sequence ς of substitutions, joins and node adds; such that *proj*$(\mathcal{Q}, \varsigma(\mathcal{D})) \neq \emptyset$.

Once a measure of semantic proximity is available, we can introduce a ranking protocol to show the user the answers in the order of most relevant to least relevant. We take as starting point the partial orders in the set of transformations defined.

Definition 14. Given a support $\mathcal{S}$, let $\mathcal{Q}$ and $\mathcal{D} = \{\mathcal{D}_i\}_{i \in I}$ be the BGs associated to a query and a documentary database, and let $\mathcal{A}^{\mathcal{D}}_{\mathcal{Q}}$ be the collection of answers obtained through a set $\mathcal{T}^{\mathcal{D}}_{\mathcal{Q}}$ of graph transformation sequences applied to get a projection of $\mathcal{Q}$ on some $\mathcal{D}_i,\ i \in I$. We then define a *ranking function associated to* $\mathcal{Q}$ *and* $\mathcal{D}$ as the ordering naturally induced in $\mathcal{A}_{\mathcal{Q}}^{\mathcal{D}}$ by any partial order on $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$.

We focus on a partial order, considering an approximate (resp. plausible) answer more relevant than a plausible (resp. partial) one. For a same type, relevance is inversely proportional to the number of transformations applied.

Definition 15. Given a support $\mathcal{S}$, let $\mathcal{Q}$, $\mathcal{D} = \{\mathcal{D}_i\}_{i \in I}$ be the BGs associated to a query and a documentary database, and let $\mathcal{A}_{\mathcal{Q}}^{\mathcal{D}}$ be the collection of answers obtained through a set $\mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ of graph transformation sequences applied on $\mathcal{Q}$ to get a projection on some $\mathcal{D}_i,\ i \in I$. We then define the *Genest's partial order on elements* $t,\ t' \in \mathcal{T}_{\mathcal{Q}}^{\mathcal{D}}$ as follows:

$$t <_G t' \ \mathit{iff} \begin{cases} t' & \textit{associates approximate answer } \text{OR} \\ t & \textit{associates a partial answer } \text{OR} \\ t \ (\textit{resp. } t') & \textit{associates a partial (resp. plausible) answer } \text{OR} \\ t,\ t' & \textit{associate the same type of answer } \text{AND} \ \mid t \mid > \mid t' \mid \end{cases}$$

while that

$$t =_G t' \ \mathit{iff} \ t \ \text{AND} \ \ t' \ \textit{associate the same type of answer } \text{AND} \ \mid t \mid = \mid t' \mid$$

We do not apply explicit document length normalization (resp. graph-based term weighting), since we assume the scale is provided by graph-ranking computation (resp. the results seem to be similar, despite its simplicity).

5. Surfing the meaning

We briefly describe the protocol for the extraction and representation of the knowledge contained in the texts [21], using BGs directly generated from them. This will allow us to exploit the IR model introduced.

5.1. *Knowledge acquisition*

This task is a chain of lexical, syntactic and semantic analysis with minimal user intervention. In relation to the former, the only requirement is on its output, that must include all possible lexical categories for a given occurrence of a form and it is denoted for description purposes as indicated below, introducing some additional structural details in order to later integrate semantic data.

Definition 16. Let $\{s_i\}_{1 \leq i \leq n}$ be the sequence of sentences in a *corpus* $\mathscr{C}$ and $\Theta_{i,j},\ 1 \leq j \leq \mid s_i \mid$ be the *occurrence* of a *form* in the i-th sentence, s_i. We denote the association of the lexical category (a) and semantic class (b) to this form, in this sentence, by $\Theta_{i,j}^{a,b}$ and we call it *term.* We use an anonymous-variable notation, $\Theta_{i,j}^{a,-}$, in order to designate the set of terms that can only be differentiated by their semantic class, which we call *token.* We denote by $\Theta_{i,j}^{-,-}$ the set of tokens referring to the same occurrence of a form, which we call *cluster.*

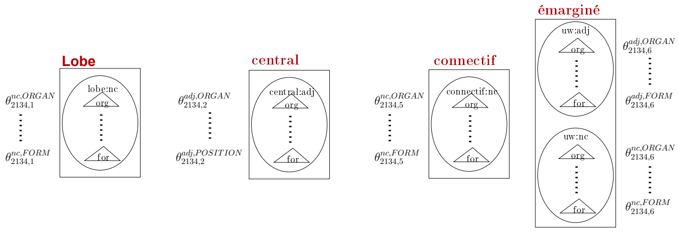

Lobe
$\theta^{nc,ORGAN}_{2134,1}$
$\theta^{nc,FORM}_{2134,1}$
lobe:nc
org
for
central
$\theta^{adj,ORGAN}_{2134,2}$
$\theta^{adj,POSITION}_{2134,2}$
central:adj
org
for
connectif
$\theta^{nc,ORGAN}_{2134,5}$
$\theta^{nc,FORM}_{2134,5}$
connectif:nc
org
for
émarginé
uw:adj
org
for
$\theta^{adj,ORGAN}_{2134,6}$
$\theta^{adj,FORM}_{2134,6}$
uw:nc
org
for
$\theta^{nc,ORGAN}_{2134,6}$
$\theta^{nc,FORM}_{2134,6}$

Fig. 2. Lexical notation

Table 1. The set $\mathscr{T}$ of initial semantic classes (types) for the *corpus* $\mathscr{B}$

Entities	**Lemmas** (in French)
organ	limbe , pétale, inflorescence, staminode, sépale, calice, corolle, foliole, rostelle, sinus, ...
fruit	fruit, samare, drupe, capsule, akène, infrutescence, gousse, ...
Properties	**Lemmas** (in French)
color	blanc, vert, rouge, bronze, argenté, brillant, clair, foncé, gris, grisâtre, noir, marron, ...
form	punctiforme, filiforme, oblongode, cunéiforme, ovale, elliptique, ové, circulaire, ...
size	minuscule, mince, large, nombreux, réduit, court, grand, gros, grêle, rétréci, égal, ...
texture	charnu, cilié, cireux, tomenteux, translucide, velu, écailleux, fibreux, glaberscent, ...
position	pendant, dorsal, axial, central, couché, amplexical, apical, basal, incombant, ...

We also consider a free-variable notation, using capital letters, in order to enumerate a range of values. So, for example, $\Theta^{a,X}_{i,j}$ refers the sequence of terms in the token $\Theta^{a,-}_{i,j}$, whose semantic class X is applicable in that context. We can naturally extend this notation to occurrences of tokens and clusters.

Table 2. A sample section from the collocations file for *corpus* $\mathscr{B}$

Word (in French)	**Position**	**Class**	**Word** (in French)	**Position**	**Class**
teinté	[2]	color	épaisseur	[1]	size
texture	[2]	texture	atteindre	[1]	organ/fruit
taille	[1]	organ/fruit	taille	[2]	size
teinte	[1]	organ/fruit	teinte	[2]	color
couleur	[1]	organ/fruit	couleur	[2]	color
texture	[1]	organ/fruit	texture	[2]	texture
forme	[1]	organ/fruit	forme	[2]	form
position	[1]	organ/fruit	position	[2]	position
altitude	[1]	organ/fruit	environ	[2]	size
tache	[1]	organ/fruit	tache	[2]	color
longueur	[1]	size	formé	[2]	organ/fruit
composé	[1,2]	organ/fruit	dépassant	[2]	size
diamètre	[1]	size	contour	[2]	form/texture
contour	[2]	form/texture	bord	[2]	forme

We illustrate these concepts, for the sentence *"lobe central du connectif émarginé"* ("emarginate central lobe of the connective") taken from *corpus* $\mathscr{B}$, in Fig. 2. Terms are here represented by triangles, tokens by ellipses and clusters by rectangles. The semantic classes associated to terms are derived either from the user or by means of a some trustworthy method. In the first case, we take the set shown in Table 1, which we organize as entities ($\mathscr{E}$) and properties ($\mathscr{P}$). In the second one, the system looks for *collocations*, sequences of words that co-occur more often than would be expected by chance and in which they keep their original meaning. We represent them by triples *marker-position-semantic class*, as shown in Table 2. The marker identifies the collocation for which the form in the indicated position can be associated with the semantic class, as in the sentence *”de couleur jaune”* ("yellow color”), where the presence of the marker *”couleur”* ("color”) reveals that *”jaune”* ("yellow”) is an instance of the semantic class "color”.

Similarly, the parse should be summarized in a graph that compiles all the possible head-dependent relationships within the text analyzed, which also requires a formal definition.

Definition 17. Let s_i, $1 \leq i \leq n$ be the i-th sentence in a *corpus* $\mathscr{C}$ and τ be the sequence of the grammar rules necessary to generate the token $\Theta^{c,-}_{i,k}$ from the token $\Theta^{a,-}_{i,j}$ in the head-dependent graph. We denote the *dependency between the tokens* $\Theta^{a,-}_{i,j}$ *and* $\Theta^{c,-}_{i,k}$, labeled by τ, as $\delta^{\theta^{a,-}_{i,j},\tau,\theta^{c,-}_{i,k}}$. The notation can be naturally extended to terms and clusters.

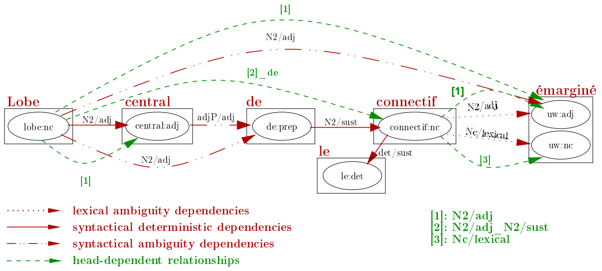


Fig. 3. Head-dependent relationships

Continuing with our example, Fig. 3 shows the head-dependent graph using dotted lines connecting the nodes involved in each case. We can observe the impact that both lexical and syntactic ambiguities have on the number of possible dependencies that go forward to the semantic analysis stage. In the first case, they multiply in relation to the number of tokens in a single cluster, or in other words, to the number of lexical categories that can be assigned to a form in a given position of a given sentence. In the second, we can see an analogue effect resulting from the multiplication of dependencies on the modifiers. An example of this would be *”émarginé”* ("emarginate”) , which could be a modifier of either *”lobe”* ("lobe”) or *”connectif”* ("connective”). This is a well-known phenomenon linked to the association of adjectival attachments to a nominal phrase, and which provides us with two possible interpretations for the sentence: "-emarginate central lobe- of the connective” and "central lobe of the -emarginate connective-”, where indents are here used to separate the syntagmas more clearly in each case.

There are still situations in which ambiguities are exclusively of semantic origin. An example is the use of prepositional structures relating multiple entities between each other, as in *"feuille avec pétiole à pubescence éparse"* ("leaf with petiole with sparse pubescence"), where *"à pubescence"* ("with pubescence") could be attached either of the nouns *"pétiole"* ("petiole") or *"feuille"* ("leaf"). Here, there is only one way to solve the problem, which is to understand the precise nature of the plant organs concerned, something that bears no relationship to the language's morphology or grammar. In any event, and regardless of its origin, an ambiguity corresponds to a situation where a dependent token has more than one head token. Such a condition provides a simple mechanism for solving non-determinism, namely to filter out the less plausible dependencies in favor of the most plausible ones, thereby ensuring that a dependent token has only one head.

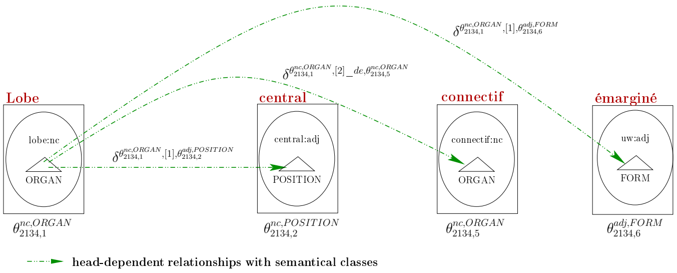

$\delta^{\theta^{nc,ORGAN}_{2134,1},[1],\theta^{adj,FORM}_{2134,6}}$
$\delta^{\theta^{nc,ORGAN}_{2134,1},[2]\_de,\theta^{nc,ORGAN}_{2134,5}}$
Lobe
central
connectif
émarginé
lobe:nc
ORGAN
$\delta^{\theta^{nc,ORGAN}_{2134,1},[1],\theta^{adj,POSITION}_{2134,2}}$
central:adj
POSITION
connectif:nc
ORGAN
uw:adj
FORM
$\theta^{nc,ORGAN}_{2134,1}$
$\theta^{nc,POSITION}_{2134,2}$
$\theta^{nc,ORGAN}_{2134,5}$
$\theta^{adj,FORM}_{2134,6}$
head-dependent relationships with semantical classes

Fig. 4. The semantic of a sentence

The prioritization of dependencies is a task for the semantic analyzer, one that it applies in three steps. The first of these is the categorization of tokens, computing the most probable token for each cluster in the text. In other words, we want to determine the lexical category for each occurrence of a given form in the position of a sentence in the *corpus*. The second step estimates the viability of dependencies between tokens, while the final categorization process seeks to attach the semantic class to the tokens involved in a given dependency. In all these three steps, the analyzer extrapolates the estimations from a local level (sentence) to a global one (*corpus*) or, in other words, initial data obtained at sentence level are combined and evaluated throughout the whole *corpus* in order to extract new conclusions that can then be applied in each sentence, the process recommencing iteratively. The user can fix an approximation threshold (resp. a maximum number of iterations) to apply in any of these procedures: υ_{to}, υ_{dto} and υ_{dte} (resp. ι_{to}, ι_{dto} and ι_{dte}). The final outcome of the reporting process is a structure that we call the *semantic of the corpus* $\mathscr{C}$ we are working with.

Definition 18. Let $\{s_i\}_{1 \leq i \leq n}$ be the sentences in a *corpus* $\mathscr{C}$, and $\mathscr{T}$ (resp. $\mathscr{F}$) be the set of semantic classes (resp. forms) associated to $\mathscr{C}$ (resp. to $\mathscr{T}$) by means of some reliable technique. We define the *semantic of the corpus* $\mathscr{C}$ as:

$$\mathscr{S}_{\mathscr{C}} := \{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}},\ P(\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}})_{\text{local}} = \max\{P(\delta^{\Theta^{X,Y}_{i,j},Z,\Theta^{V,W}_{i,k}})_{\text{local}}\}\} \tag{2}$$

where *max* is the maximal function on $\mathbb{N}$, P_{local} computes the probability for its argument at local level, and $\delta^{\Theta^{X,Y}_{i,j},Z,\Theta^{V,W}_{i,k}}$ are the dependencies computed as a result of the knowledge acquisition process previously described. We can naturally restrict the concept to refer the *semantic of a document $\mathscr{D}$ in $\mathscr{C}$* (resp. *of a sentence $s_i \in \mathscr{D}$*) by

$$\mathscr{S}^{\mathscr{D}}_{\mathscr{C}} := \{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{C}},\ s_i \in \mathscr{D}\} \quad \text{(resp. by } \mathscr{S}^{\mathscr{D},s_i}_{\mathscr{C}} := \{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}^{\mathscr{D}}_{\mathscr{C}}\}) \tag{3}$$

Returning to our case scenario example, Fig. 4 shows the result of the knowledge acquisition task for the same sentence whose initial graph of dependencies was shown previously in Fig. 3. The comparison between the two figures highlights the magnitude of the simplifications that have been made.

5.2. *Knowledge representation*

We are now ready to structure the BGs that shape the meaning of a text. Although the proposal is independent of the knowledge domain considered, it is necessary to locate our work in a concrete one, in order to suitably model the support serving as a basis for subsequently defining such graphs. As already justified, our choice is the botanical description from the *corpus $\mathscr{B}$*. In this sense, we retake the set $\mathscr{T}$ of semantic classes (types) shown in Table 1, in order to introduce a partial order on it as follows:

$$\forall\, t \in \mathscr{E} = \{\textit{fruit, organe}\},\ t \leq \varepsilon \leq \top \tag{4}$$

$$\forall\, t \in \mathscr{P} = \{\textit{couleur, forme, taille, texture, position}\},\ t \leq \rho \leq \top \tag{5}$$

where ε (resp. ρ) is the greater element for the entities (resp. properties) $\mathscr{E}$ (resp. $\mathscr{P}$). In this way, we introduce our running support $\mathcal{S}_{\mathscr{B}} = (\mathcal{T}_{\mathcal{C}_{\mathscr{B}}}, \mathcal{T}_{\mathcal{R}_{\mathscr{B}}}, \mathcal{I}_{\mathscr{B}})$ by defining:

$$\mathcal{T}_{\mathcal{C}_{\mathscr{B}}} := \{\varepsilon, \rho\} \cup \mathscr{E} \cup \mathscr{P} \cup \{\top\} \tag{6}$$

$$\begin{aligned}\mathcal{T}_{\mathcal{R}_{\mathscr{B}}} := \{&[b,\tau,d], [b,*,d],\ \exists\, \delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{B}}\} \cup \{[\varepsilon,*,\varepsilon]\} \cup \{[\varepsilon,*,\rho]\} \cup \{[\rho,*,\rho]\cup \\ &\{[\top,*,\top]\}\end{aligned} \tag{7}$$

$$\mathcal{I}_{\mathscr{B}} := \{\Theta^{a,-}_{i,j},\ \Theta^{c,-}_{i,k}\}_{\delta^{\Theta^{a,-}_{i,j},-,\Theta^{c,-}_{i,k}}} \tag{8}$$

where $\mathscr{S}_{\mathscr{B}}$ is the semantic associated with the running *corpus $\mathscr{B}$*. With regard to the set of relations $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$, these are directly extracted from $\mathscr{S}_{\mathscr{B}}$ through the transitional dynamic, and summarize a transition between two terms from the point of view of the semantic classes (types) involved. As extra elements, we add triples representing any possible transition in the semantically related generic concepts. The partial order we consider in $\mathcal{T}_{\mathcal{C}_{\mathscr{B}}}$ (resp. $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$) is naturally induced by the one previously defined in $\mathscr{T}$ (resp. $\mathscr{T}$ and $\mathcal{I}_{\mathscr{B}}$). Finally, we define the markers $\mathcal{I}_{\mathscr{B}}$ as the set of forms in the *corpus $\mathscr{B}$*.

Our starting point to introduce the BGs on this support is the semantic $\mathscr{S}_{\mathscr{D}_m}$ associated with each of the M documents in the *corpus* $\mathscr{B} = \bigcup_{m \in M} \mathscr{D}_m$, as follows:

$$\mathcal{C}_{\mathscr{D}_m} := \{\Theta^{a,b}_{i,j}, \Theta^{c,d}_{i,k}\}_{\delta^{\Theta^{a,b}_{i,j},\_,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{D}_m}} \qquad \mathcal{R}_{\mathscr{D}_m} := \{[b,\tau,d],\ \exists\ \delta^{\Theta^{\_,b}_{\_,\_},\tau,\Theta^{\_,d}_{\_,\_}} \in \mathscr{S}_{\mathscr{D}_m}\} \tag{9}$$

$$\mathcal{E}_{\mathscr{D}_m} := \bigcup_{\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}} \in \mathscr{S}_{\mathscr{D}_m}} \{([b,\tau,d],1,\Theta^{a,b}_{i,j}),([b,\tau,d],2,\Theta^{c,d}_{i,k})\} \tag{10}$$

$$\mathcal{L}_{\mathscr{D}_m}(X) := \begin{cases} [b,\Theta^{a,\_}_{i,j}] & \text{if } X = \Theta^{a,b}_{i,j} \in \mathcal{C}_{\mathscr{D}_m} \\ X & \text{if } X \in \mathcal{R}_{\mathscr{D}_m} \\ 1 & \text{if } X = (\_,1,\_) \in \mathcal{E}_{\mathscr{D}_m} \\ 2 & \text{if } X = (\_,2,\_) \in \mathcal{E}_{\mathscr{D}_m} \end{cases} \tag{11}$$

Succinctly, a conceptual node in $\mathcal{C}_{\mathscr{D}_m}$ is any term involved in the semantic $\mathscr{S}_{\mathscr{D}_m}$, while relation nodes in $\mathcal{R}_{\mathscr{D}_m}$ are elements of $\mathcal{T}_{\mathcal{R}_{\mathscr{B}}}$ associated to transitions in $\mathscr{S}_{\mathscr{D}_m}$. The multiset of edges $\mathcal{E}_{\mathscr{D}_m}$ contains in this case only binary relations, the head (resp. dependent) term corresponding to the first (resp. second) triple. With regard to the labeling function $\mathcal{L}_{\mathscr{D}_m}$, it makes it possible to recover the class and the token associated to a given term representing a concept, whilst implementing the identity on the relations, since in our case we build these directly from $\mathscr{S}_{\mathscr{D}_m}$. The value of this function on edges identifies head (1) and dependent edges (2). In order to cushion this notational load, we introduce a simplified graphic representation for BGs, much more intuitive and visual, translating the transitional dynamic from the parse directly to the graph.

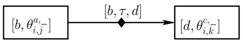


Fig. 5. A representation as BG for a dependency $\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}}$

Let's assume a dependency $\delta^{\Theta^{a,b}_{i,j},\tau,\Theta^{c,d}_{i,k}}$ in the parse. This involves the head (resp. the dependent) concept $\Theta^{a,b}_{i,j}$ (resp. $\Theta^{c,d}_{i,k}$), a relation $[b,\tau,d]$ as well as the corresponding edge $([b,\tau,d],1,\Theta^{a,b}_{i,j})$ (resp. $([b,\tau,d],2,\Theta^{c,d}_{i,k})$); we summarize in the graph shown in Fig. 5. As a practical example, we can see in Fig. 6 the BG corresponding to the sentence whose graph of head-dependent relationships and associated semantic were shown in Figs. 3 and 4, respectively. To provide an overall understanding we do not make explicit the indexes corresponding to either the number of the sentence in the text or the position of the form in that sentence.

6. Using the tool

The following is an informal description of our system, hereinafter referred to as *COnceptual and General Information Retrieval* (COGIR), at work. In order to complement the core functionality described above, we developed a graphical user in-

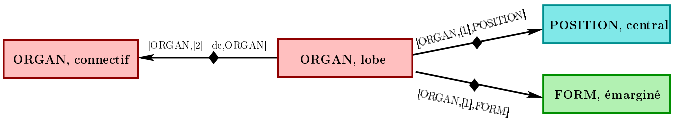


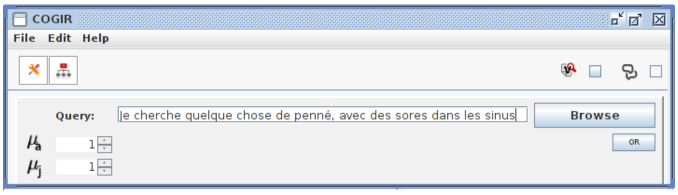
Fig. 6. The BG of a sentence

terface featuring a *corpus* management component, configuration dialogs as well as graph-based operations and visualizations.

6.1. *The interface*

The general appearance of the interface can be seen in Fig. 7, which reflects the state of the *main panel* for a concrete input query, which we can enter through the dialog box included to that effect. Just above of this, we find two buttons to the left, the first of which serves to select the linguistic and semantic resources for the IR task[c]. The second button enables us to open two extra panels as shown in Fig. 8 for the Acridocarpus_camerounensis, the left-most of which we baptize as *documentary panel* and shows the list of documents in the corpus.

We can select a document, whose conceptual representation is then visualised to the right in the *conceptual panel*, omitting in part the description of relations to avoid saturating the visualization task, although such data is considered internally. Thus, lexical information associated with the label relation is kept if present, while syntactic information is withheld. This is the case of the dependence close to the legend at bottom left, between the concepts [ORGAN,calice] and [ORGAN,sépale], whose relation should be [ORGAN,[2]_à,ORGAN] but is nevertheless labeled only by ` a.



Query: Je cherche quelque chose de penné, avec des sores dans les sinus

μ_a 1

μ_j 1

Fig. 7. The main panel of COGIR

Continuing with the interface description, at the top right of the main panel, we find two check-boxes. The first of these ☑ serves to activate the visual querying mode, whilst the second ☑, useful for relevance feedback purposes, allows us to

[c] we include here, for example, the access to the documentary collection we want to study, as well as the initial set of semantic classes we consider in the domain to initialize the knowledge acquisition process.

indicate that the set of texts in which we are looking for an answer is retaken from the list of documents recovered from the previous query. In the lower-left part of the main panel, we can distinguish another two smaller dialog boxes for entering the values for the parameters μ_a and μ_j limiting the number of adds and joins on the query graphs during the search task. Following the main dialog box to the right of the same panel, we have a button Browse that allows us to raise the query against the *corpus*. Just below this, another button OR enables the user to pose a finite number of queries simultaneously if he should so wish.

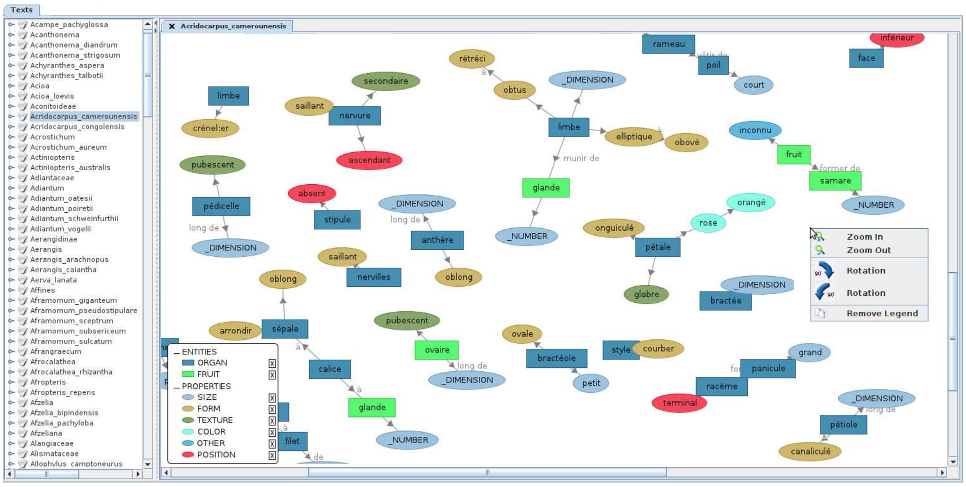


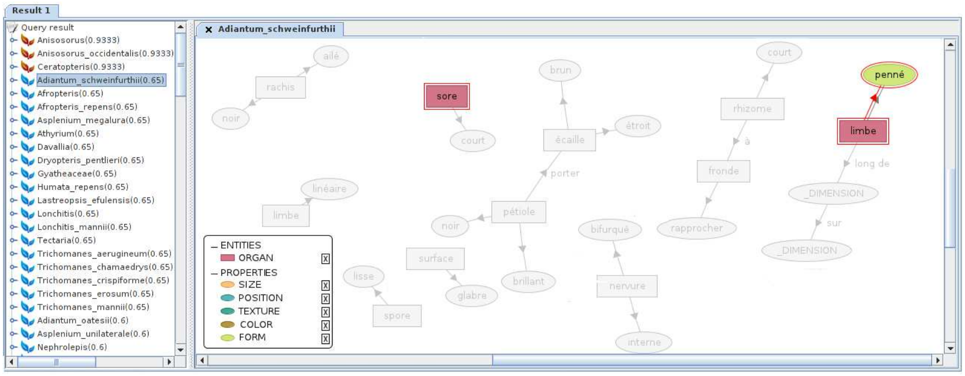
Fig. 8. The documentary and conceptual panels of COGIR, on the database

Having raised the query, both the documentary and the conceptual panels are updated, as we can see in Fig. 9. The first then shows the list of documents for which some kind of conceptual relation with the query has been detected. Using the same font color, this list pools together those documents associated to a same type of answer: approximate, plausible or partial. In our example, the query has only produced two kinds of answer: the first three are plausible, while the rest are partial ones. The user can then select any one of these documents to visualize in the conceptual panel the semantic representation as well as, using a darker background color, the nodes and relations involved in the projection from the query to this particular text, which here corresponds to the entry Adiamtum_schweinfurthii. This allows us to obtain a precise contextualization of the search task, facilitating a more comprehensive start of a possible relevance feedback loop from the set of documents retrieved and a new query. To do so, the user should indicate the corresponding option in the check-box ☑ at the upper-right of the main panel.

The interface also makes it possible to question the system excusively using a visual protocol. On any conceptual representation, obtained either from a document in the *corpus* as in Fig. 8 or from a response to a query as in Fig. 9, the user can directly mark the set of relations between concepts he is looking for in the database with the mouse and send them immediately to the search engine using the same button Browse as we had earlier done in the case of text queries. Compared to the latter, graphical queries allow the user to work with explicit semantic representations, which can be of interest in order to introduce greater precision and reliability in the search.

6.2. *The system at work*

Once we have sketched the interface, we can learn a little more about how and why a response is built in practice. To do so, we illustrate the mechanisms of conceptual IR applied to SB on the basis of a practical example, the query *”Je cherche quelque chose de penné, avec des sores dans les sinus”* ("I am looking for something pinnate, with sores in the sinuses”) that we previously entered through the dialog box in the main panel, as shown in Fig. 7. This query involves both taxonomic and non-taxonomic relationships as well as attributes, a common component in classic ontology theory [22]. In the first case, we refer *”quelque chose avec des sores”* ("something with sores”), which we can formalize by "sores part-of something”. In the second, we include a locative relationship taken from PO, *”sores dans les sinus”* ("sores in the sinuses”), corresponding to an annotation of the type "sores located-in sinuses”. In the case of attributes, we find *”quelque chose de penné”* ("something pinnate”), which we can characterize by "something has-property pinnate”. It is important to remember here that PO, the ontological reference in the botanic domain, limits the use of attributes to aspects related to synonym and references, generally covering terms[d] and not phenotypes. This justifies the need for more specific ontologies, such as PATO[e] [31].



Query result
Anisosorus(0.9333)
Anisosorus_occidentalis(0.9333)
Ceratopteris(0.9333)
Adiantum_schweinfurthii(0.65)
Afropteris(0.65)
Afropteris_repens(0.65)
Asplenium_megalura(0.65)
Athyrium(0.65)
Davallia(0.65)
Dryopteris_pentlieri(0.65)
Gyatheaceae(0.65)
Humata_repens(0.65)
Lastreopsis_efulensis(0.65)
Lonchitis(0.65)
Lonchitis_mannii(0.65)
Tectaria(0.65)
Trichomanes_aerugineum(0.65)
Trichomanes_chamaedrys(0.65)
Trichomanes_crispiforme(0.65)
Trichomanes_erosum(0.65)
Trichomanes_mannii(0.65)
Adiantum_oatesii(0.6)
Asplenium_unilaterale(0.6)
Nephrolepis(0.6)

Fig. 9. The documentary and conceptual panels of COGIR, on a query's answer

[d] which we can identify with our entities.
[e] see http://obofoundry.org/wiki/index.php/PATO:Main_Page

Among the set of answers returned by the system, we focus on the partial one already seen in Fig. 9. We can observe the existence of three conceptual nodes, [ORGAN, sore] to the left, and [ORGAN,limbe] and [FORM,penné] to the right. The latter two are linked by a relation node [ORGAN,[1],ORGAN] that, as previously mentioned, we do not visualize in the interest of understanding. To obtain this answer, a chain of transformations has been applied in order to build a projection from the conceptual representation of the query in Fig. 10 to the document:

(1) The addition of the node concept [ORGAN,sinus].
(2) The addition of the relation [ORGAN,[2]_dans,ORGAN], between the concepts [ORGAN, sore] and [ORGAN,sinus].
(3) The addition of the relation [ORGAN,[2]_avec,ORGAN], between the concepts [ORGAN, limbe] and [ORGAN,sore].

The result is the BG ilustrated in Fig. 11, which subsumes the one shown in Fig. 10. More exactly, [ORGAN,limbe] is a concept node specializing [ε,*], which in turn represents *"quelque chose"* ("something"). The same applies to the relation nodes linking [ε,*] with [ORGAN,sore] and [FORM,penné], since in both cases they can be specialized in [ORGAN,[2]_avec,ORGAN] and [ORGAN,[1],FORM], respectively.

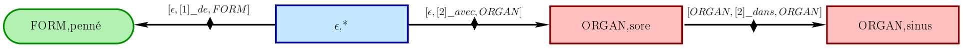


Fig. 10. The BG of the query

Visual querying is performed from a conceptual representation that, as said previously, can be associated to a database or to the trace of a projection. In both cases the protocol to follow is the same, which we illustrate in the context of this last scenario in order to provide relevance feedback. The goal is to identify new contents among those texts containing some answer to the original query, which implies searching in the list of the documentary panel in Fig. 9, taking as our starting point a set of concepts and relations we combine from the current projection. Once we have activated the two check-boxes in the upper-right part of the main panel, enabling both the visual querying ☑ and the relevance feedback ☑ modes, the new query is built using the mouse to select the concepts involved in the new search and to define new relations between them.

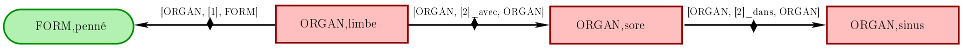
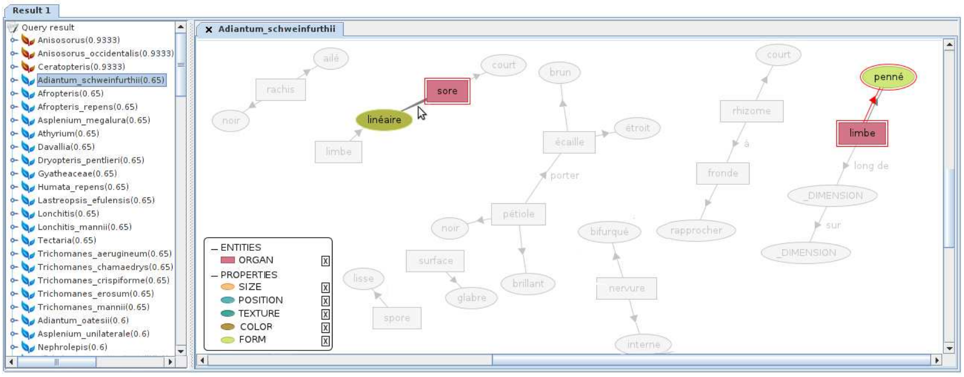


Fig. 11. Resulting BG from a chain of transformations

As a case in point, let us assume that we want to find those documents including an additional reference to *"sore linéaire"* ("linear sore"), which involves the concepts [ORGAN,sore] and [FORM,linéaire], corresponding respectively to the head and the dependent of a new link we must define. To do this, the user first select the future head for latter establishing the dependency simply by dragging with the mouse to the dependent. In order to distinguish links of this kind from those associated to the original BG, its graphical representation is slightly different, as we can see in Fig. 12. We are now ready to raise the new search process, simply by pressing the button Browse, to obtain the set of answers shown in the documentary panel of Fig. 13. In contrast to the situation in the original query, the system now covers the whole spectrum of answers, including five approximate ones in the highest position on the list. Likewise, the conceptual panel now shows the concepts and relations involved in the projection associated to the visual query with respect to the document Loxograma_latifolia. The picture allows us not only to contextualize the search process, but also to appreciate that the answer is an exact one since no transformations are needed to build the projection from the conceptual representation of the query in the conceptual panel of Fig. 12 to the text one in Fig. 13.



Result 1

Query result
Anisosorus(0.9333)
Anisosorus_occidentalis(0.9333)
Ceratopteris(0.9333)
Adiantum_schweinfurthii(0.65)
Afropteris(0.65)
Afropteris_repens(0.65)
Asplenium_megalura(0.65)
Athyrium(0.65)
Davallia(0.65)
Dryopteris_pentlieri(0.65)
Gyatheaceae(0.65)
Humata_repens(0.65)
Lastreopsis_efulensis(0.65)
Lonchitis(0.65)
Lonchitis_mannii(0.65)
Tectaria(0.65)
Trichomanes_aerugineum(0.65)
Trichomanes_chamaedrys(0.65)
Trichomanes_crispiforme(0.65)
Trichomanes_erosum(0.65)
Trichomanes_mannii(0.65)
Adiantum_oatesii(0.6)
Asplenium_unilaterale(0.6)
Nephrolepis(0.6)

Fig. 12. An example of visual query

7. The Performance

Our aim now is to provide a simple and comprehensible overview of the performance achieved by our IR architecture. To this end, we first need to define a formal testing frame in order to guarantee the relevance of the conclusions drawn from the experimental evidence.

7.1. *The testing frame*

In order to evaluate COGIR we need a representative documentary collection, a set of topics, a set of trustworthy evaluation measures and a baseline serving as referal. In the former case, we choose the *corpus* $\mathscr{B}$ for the reasons previously stated. To select the topics, we take up the proposal described in [21], which includes 150 queries distributed equitably among three levels of difficulty (low, medium and high) and formally motivated. Under the heading of evaluation metrics, we differentiate two groups of measures: set-based and rank-based. The former focuses on the relevant or non-relevant character of the documents retrieved, including *precision* and *recall*, as well as F and *fall-out* measures. These two latter allow us to estimate the harmonic mean of precision and recall and to take into account the proportion of non-relevant documents retrieved, respectively. The second group also takes into account the order in which the returned documents are presented. We here consider *precision* (resp. *recall*) *at k documents* (P@k) (resp. R@k), which permits us to compute these parameters even when we are only interested in fixed low levels of retrieved results as is typically the case of a web search. To this respect we consider $k = 10$, corresponding approximately to the size of the first page of results returned by a search engine, which seems to identify the self-focusing threshold for users [25]. The geometric interpretation of the precision-recall graph corresponds to the *mean average precision* (MAP). In order to highlight improvements for low-performing queries, we calculate the *geometric mean average precision* (GMAP). With the intention of distinguishing between documents that are explicitly judged as non-relevant and those that are only assumed to be non-relevant because they are unjudged, we also consider the *binary preference relation* (BPREF). We also include the *normalized discounted cumulative gain* (NDCG) for the purpose of separately evaluating the performance at each relevance level, penalizing the fact that highly relevant documents appear lower in a search result. Finally, for comparative purposes we consider the following ranking functions:

(1) As algebraic distances, pivoted cosine [48] and impact-based ranking [2], both using a weighting factor TF-IDF [45]. The *slope* in the former was tuned from 0 to 0.44 in increments of 0.04, finally taking 0.44.
(2) As probabilistic ranking, Okapi's BM25 [32]. We tuned the b parameter from 0 to 1 in increments of 0.05, obtaining $b = 0.3$. With regard to k_1 and k_3, it seems that they have little effect on retrieval performance[f], so we fixed them to 1.2 and 1,000 respectively, as indicated by [46].
(3) As language model measure, Dirichlet Smoothing [57]. We tune the μ parameter from 1,000 to 3,000 in increments of 100, resulting in $\mu = 2,800$.

We chose ZETTAIR[g] as the common implementation platform for all these metrics. As both ZETTAIR and COGIR are written in C, this allows us to minimize the impact of implementation features on the tests.

As a general parameter setting for our IR proposal, we tune the ratios μ_j and μ_a from 0 to 0.5 in increments of 0.05, obtaining respectively the values 0.2 and 0.3. In this respect, as above, all parameters were tuned taking the MAP as reference.

[f] see `http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=baselines`
[g] see `http://www.seg.rmit.edu.au/zettair/`

As thresholds for the categorization processes we take $\upsilon_{to} = 0.7$, $\upsilon_{dto} = 0.7$ and $\upsilon_{dte} = 0.8$. We fix the maximum number of iterations in all cases (ι_{to}, ι_{dto} and ι_{dte}) to 10.

7.2. *The experimental results*

We can now input, visualize and interpret the results according to the different evaluation metrics, which we have condensed in Tables 3 and 4 according to whether they take into account the order in which the returned documents are presented, or not, respectively. We use bold (resp. underlined) fonts to mark the best overall values (resp. the baselines). Each value associates in brackets the percentage of improvement with regard to the corresponding baseline, reporting its statistical significance with respect to the latter ($p < 0.05$) using the Wilcoxon matched-pairs signed-ranks test [54] and marking it with a star.

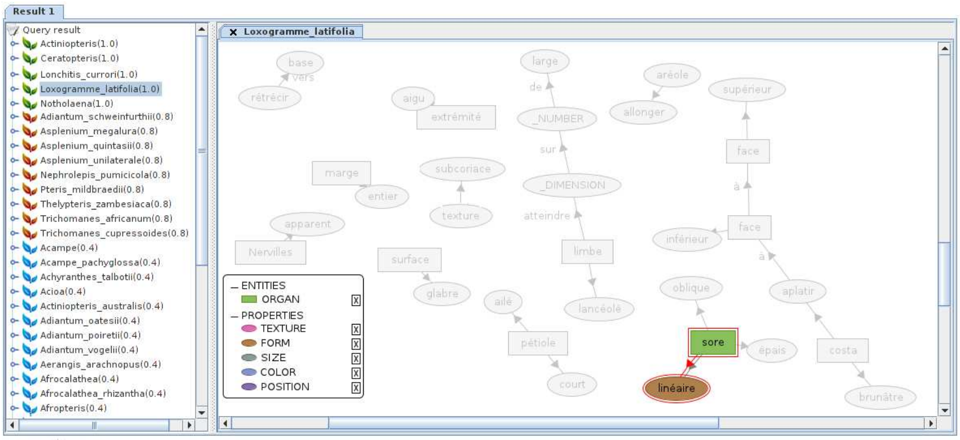


Fig. 13. The relevance feedback applied on the BG of the visual query

Set-based evaluation favors COGIR over the rest of IR systems on all metrics. With respect to the other ranking models, all of them produce similar results, with some differences that seem irrelevant when set against those previously mentioned for the conceptual approach[h]. The best (resp. the worst) absolute increase in the percentage for all ranking metrics with COGIR is reached for precision (resp. recall) (108.53%) (resp. 37.14%), while regarding numerical values, the best (resp. the worst) one is reached by recall (resp. F-measure) (0.6096) (resp. 0.2221). We need to remember that the fall-out is a negative measure, in the sense that the best

[h] the minimum increase for COGIR in relation to the baseline corresponds to recall with 37.14%, while the minimum decrease for the rest of ranking models is associated to BM25 on F-measure with -0.16%.

results are associated to minimum values. On the whole, these data allow us to argue the importance of including mechanisms for an efficient semantic treatment on SB.

Table 3. Results on set-based evaluation measures

	Precision	Recall	F-measure	Fall-out
COGIR	**0.3495** (108.53%)*	**0.6096** (37.14%)*	**0.2221** (82.49%)*	**0.1100** (-57.85%)*
BM25	0.1676	0.4427 (-0.40%)	0.1215 (-0.16%)	0.2908 (11.41%)
DIRICHLET	0.1535 (-8.41%)	0.4055 (-8.77%)	0.1113 (-8.54%)	0.2908 (11.41%)
IMPACT	0.1676	0.4445	0.1217	0.2610
PIV. COSINE	0.1669 (-0.41%)	0.4409 (-0.81%)	0.1210 (-0.57%)	0.2908 (11.41%)

Ranked-based evaluation corroborates the set-based one, supporting the idea that the conceptual strategy better exploits the semantic relations which make up the meaning of the text[i]. The best (resp. the worst) absolute increase in the percentage for all ranking functions in the case of COGIR is reached for the MAP (resp. the P@10) (124.38%) (resp. 51.73%) metric, while regarding numerical values the best (resp. the worst) one is reached by NDCG (resp. GMAP) (0.7317) (resp. 0.3480). All of the ranking functions achieve their best numerical values for the NDCG measure, which suggests that documents are successfully evaluated at each relevance level. Curiously, all the minimum numerical values also concur on GMAP, revealing the complexity of highlighting improvements for low-performing topics.

Table 4. Results on rank-based evaluation measures

	MAP	GMAP	BPREF
COGIR	**0.5088** (124.38%)*	**0.3480** (101.97%)*	**0.5063** (100.27%)*
BM25	0.2213 (-2.38%)	0.1669 (-3.13%)	0.2422 (-4.19%)
DIRICHLET	0.2080 (-8.24%)	0.1488 (-13.63%)	0.2372 (-6.17%)
IMPACT	0.1873 (-17.38%)	0.0712 (-58.67%)	0.1850 (-26.82%)
PIV. COSINE	0.2267	0.1723	0.2528
	P@10	R@10	NDCG
COGIR	**0.5432** (51.73%)*	**0.3900** (59.77%)*	**0.7317** (72.08%)*
BM25	0.3253 (-9.13%)	0.2283 (-6.47%)	0.4198 (-1.27%)
DIRICHLET	0.3207 (-10.41%)	0.2308 (-5.44%)	0.3952 (-7.05%)
IMPACT	0.2407 (-32.76%)	0.1726 (-29.29%)	0.3628 (-14.67%)
PIV. COSINE	0.3580	0.2441	0.4252

8. Conclusions

Scientific text constitutes the main source of relevant biological knowledge for researchers, although its heterogeneity and different origins deeply complicate both the automatic extraction of relevant information and the subsequent search for interactions. The problem is compounded by the avalanche of data generated and is

[i] the minimum increase for COGIR in relation to the baseline corresponds to GMAP with 34.80%, while the minimum decrease for the rest of ranking models is associated to BM25 on NDCG with -1.27%.

exacerbated when it comes to an inexpert biologist and/or computer user. With this aim, we have introduced an architecture able to give the user everything he needs in order to relate data: tools for knowledge acquisition and representation as well as for conceptual-based IR and efficient querying, all integrated into a user-friendly interface. This makes it possible to design a style of work based in the principle of *what you see is what you get*, allowing the automatic linking, transformation, overlaying, and comparison of information. In particular, it is possible to contextualize the search task, allowing the user to easily trace and understand both the IR process and the evolution of the knowledge database with respect to the changes in the underlying corpus, which in turn helps him to select contents or even to disambiguate the relationships described in the text.

Our proposal has been developed from a computational model providing soundness and completeness with regard to deduction in FOL, which guarantees a well-founded treatment of semantics. Compared to other works, the concept of ontology we apply is more flexible, moving away from complex static structures that are difficult to generate and even harder to maintain. Instead, we opt for a more dynamic approach where the relations between concepts and the concepts themselves are not subject to rigid syntactic norms, thereby facilitating their automatic generation from the text. The formal basis is the notion of CG, allowing us to avoid the problems associated with the lack of compatibility between different ontological descriptive formalisms, which notably favors the integration of data from different sources. This choice also makes it possible to solve queries when only partial knowledge is available and to capture the strongest semantic evidence that results in the most accurate similarity assessment, when dealing with overlapping knowledge. In practice, trials show promising results, which allows us to argue the viability of NLP as a basis for implementing SB.

Acknowledgments

This research has been partially funded by the Spanish Ministry of Education and Science and FEDER through project TIN2010-18552-C03-01, and by the Autonomous Government of Galicia through project CN 2012/317 and CN 2012/319.

References

1. M. Amami, R. Faiz, and A. Elkhlifi. A framework for biological event extraction from text. In *Proc. of the 2nd Int. Conf. on Web Intelligence, Mining and Semantics*, WIMS'12, pages 52:1–52:9, New York, NY, USA, 2012. ACM.
2. V.N. Anh and A. Moffat. Impact transformation: effective and efficient web retrieval. In *Proc. of the 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR'02, pages 3–10, New York, NY, USA, 2002. ACM.
3. E. Antezana, W. Blondé, A. Venkatesan, B. De Baets, V. Mironov, and M. Kuiper. Semantic systems biology: enabling integrative biology via semantic web technologies. In *Proc. of the Int. Conf. on Web Intelligence, Mining and Semantics*, WIMS'11, pages 58:1–58:5, New York, NY, USA, 2011. ACM.

4. A.A. Barforush and A. Rahnama. Ontology learning: revisted. *Journal of Web Engineering*, 11(4):269–289, December 2012.
5. E. Beisswanger, S. Schulz, H. Stenzhorn, and U. Hahn. BioTop: An upper domain ontology for the life sciences: A description of its current structure, contents and interfaces to OBO ontologies. *Applied Ontology*, 3(4):205–212, December 2008.
6. J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. Extracting complex biological events with rich graph-based feature sets. In *Proc. of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP'09, pages 10–18, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
7. O. Bodenreider and R. Stevens. Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274, 2006.
8. T.H. Cao and A.H. Mai. Ontology-based understanding of natural language queries using nested conceptual graphs. In *Proc. of the 18th Int. Conf. on Conceptual Structures: from Information to Intelligence*, ICCS'10, pages 70–83, Berlin, Heidelberg, 2010. Springer-Verlag.
9. F. Carbayo and A Marques. The costs of describing the entire animal kingdom. *Trends in Ecology & Evolution*, 26(4):154–155, 2011.
10. J.J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson. Jena: implementing the semantic web recommendations. In *Proc. of the 13th Int. World Wide Web Conf. on Alternate Track Papers & Posters*, WWW Alt.'04, pages 74–83, New York, NY, USA, 2004. ACM.
11. M. Chein and M.-L. Mugnier. *Graph-Based Knowledge Representation: Computational Foundations of Conceptual Graphs.* Advanced Information and Knowledge Processing. Springer, 2008.
12. D. Corbett. Graph-based representation and reasoning for ontologies. In J. Fulcher and L. Jain, editors, *Computational Intelligence: A Compendium*, volume 115 of *Studies in Computational Intelligence*, pages 351–379. Springer Berlin / Heidelberg, 2008.
13. H. Cui. Competency evaluation of plant character ontologies against domain literature. *Journal of the American Society for Information Science and Technology*, 61(6):1144–1165, June 2010.
14. H. Cui. Charaparser for fine-grained semantic annotation of organism morphological descriptions. *Journal of the American Society for Information Science and Technology*, 63(4):738–754, April 2012.
15. W.M. Dahdul, J.G. Lundberg, P.E. Midford, J.P. Balhoff, H. Lapp, T.J. Vision, M.A. Haendel, M. Westerfield, and P.M. Mabee. The Teleost Anatomy Ontology: Anatomical Representation for the Genomics Age. *Systematic Biology*, 59(4):369–383, 2010.
16. J. Daltio and Claudia B. Medeiros. Aondê: An ontology web service for interoperability across biodiversity applications. *Information Systems*, 33(7-8):724–753, November 2008.
17. E. de la Clergerie. Building factorized TAGs with meta-grammars. In *Proc. of 10th Int. Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+10)*, pages 111–118, New Haven, CO, United States, 2010.
18. A.R. Deans, M.J. Yoder, and J.P. Balhoff. Time to change how we describe biodiversity. *Trends in Ecology & Evolution*, 27(2):78–84, February 2012.
19. L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: a search and metadata engine for the semantic web. In *Proc. of the Thirteenth ACM Int. Conf. on Information and Knowledge Management*, CIKM'04, pages 652–659, New York, NY, USA, 2004. ACM.
20. J. Euzenat and P. Shvaiko. *Ontology Matching.* Springer Pub. Co., Inc., Secaucus, NJ,

USA, 1st edition, 2010.
21. M. Fernández. *Adquisición y representación del conocimiento mediante procesamiento del lenguaje natural.* PhD thesis, University of A Coruña, Dept. of Computer Science, A Coruña, Spain, 2012.
22. S. Omerovic G. Jakus, V.Milutinovic and S. Tomazic. *Concepts, Ontologies, and Knowledge Representation.* Springer Publishing Company, Inc., 2013.
23. D. Genest and M. Chein. A content-search information retrieval process based on conceptual graphs. *Knowledge and Information Systems*, 8(3):292–309, 2005.
24. S.E. Gordon. Eliciting and representing biology knowledge with conceptual graph structures. In K.M. Fisher and M.R. Kibby, editors, *Knowledge Acquisition, Organization, and Use in Biology*, volume 148 of *NATO ASI Series*, pages 206–225. Springer Berlin Heidelberg, 1996.
25. L.A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In *Proc. of the 27th Int. Conf. on Research and Development in Information Retrieval*, SIGIR'04, pages 478–479, New York, NY, USA, 2004. ACM.
26. B. Haddow and M. Matthews. The extraction of enriched protein-protein interactions from biomedical text. In *Proc. of the Workshop on Biological, Translational, and Clinical Language Processing*, BioNLP'07, pages 145–152, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.
27. S. Hartmann, H. Köhler, and J. Wang. Ontology consolidation in bioinformatics. In *Proc. of the Seventh Asia-Pacific Conf. on Conceptual Modelling - Volume 110*, APCCM'10, pages 15–22, Darlinghurst, Australia, Australia, 2010. Australian Computer Society, Inc.
28. S. Hensman. Construction of conceptual graph representation of texts. In *Proc. of the Student Research Workshop at HLT-NAACL 2004*, HLT-SRWS'04, pages 49–54, Stroudsburg, PA, USA, 2004. ACL.
29. S. Hensman and J. Dunnion. Constructing conceptual graphs using linguistic resources. In *Proc. of the 4th WSEAS Int. Conf. on Telecommunications and Informatics*, TELE-INFO'05, pages 34:1–34:6, Stevens Point, Wisconsin, USA, 2005. World Scientific and Engineering Academy and Society (WSEAS).
30. P. Jaiswal, S. Avraham, K. Ilic, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, D. Ware, and F. Zapata. Plant ontology (po): a controlled vocabulary of plant structures and growth stages: Research articles. *Comparative and Functional Genomics*, 6(7-8):388–397, October 2005.
31. A. Janning and H. Cui. Evaluating the botanical coverage of PATO using an unsupervised learning algorithm. In *Proc. of the 2012 iConference*, iConference'12, pages 504–505, New York, NY, USA, 2012. ACM.
32. K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management*, 36(6):779–808, November 2000.
33. A.K. Joshi. Properties of formal grammars with mixed types of rules and their linguistic relevance. In *Proc. of the Third Int. Conf. on Computational linguistics*, COLING'69, pages 1–18, Stroudsburg, PA, USA, 1969. Association for Computational Linguistics.
34. S.S. Kamaruddin, A.R. Hamdan, A.A.Bakar, and F. Mat Nor. Conceptual graph interchange format for mining financial statements. In *Proc. of the 4th Int. Conf. on Rough Sets and Knowledge Technology*, RSKT'09, pages 579–586, Berlin, Heidelberg, 2009. Springer-Verlag.
35. H. Kitano. Systems biology: a brief overview. *Science*, 295:1662–1664, 2002.

36. J.A. Legaz, M.C. Miñarro, M. Madrid, S. Torres, and J.T. Fernández. Using ontologies for supporting genomic sequence annotation projects. In *Proc. of the 2nd ACM Conf. on Bioinformatics, Computational Biology and Biomedicine*, BCB'11, pages 617–625, New York, NY, USA, 2011. ACM.
37. J. Liang, T. Nguyen, K. Koperski, and G. Marchisio. Ontology-based natural language query processing for the biological domain. In *Proc. of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, LNLBioNLP'06, pages 9–16, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
38. M. Montes, A.F. Gelbukh, and A. López. Text mining at detail level using conceptual graphs. In *Proc. of the 10th Int. Conf. on Conceptual Structures: Integration and Interfaces*, ICCS'02, pages 122–136, London, UK, UK, 2002. Springer-Verlag.
39. H. Mousselly-Sergieh and R. Unland. IROM: information retrieval-based ontology matching. In *Proc. of the 5th Int. Conf. on Semantic and Digital Media Technologies*, SAMT'10, pages 127–142, Berlin, Heidelberg, 2011. Springer-Verlag.
40. C.J. Mungall. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, 5(6-7):509–520, August 2004.
41. V. O'Day, A. Adler, A. Kuchinsky, and A. Bouch. When worlds collide: molecular biology as interdisciplinary collaboration. In *Proc. of the Seventh European Conf. on Computer Supported Cooperative Work*, ECSCW'01, pages 399–418, Norwell, MA, USA, 2001. Kluwer Academic Publishers.
42. J. Padial, A. Miralles, I. De la Riva, and M. Vences. The integrative future of taxonomy. *Frontiers in Zoology*, 7(1):1–16, 2010.
43. P. Parapatics and M. Dittenbach. Patent claim decomposition for improved information extraction. In *Proc. of the 2nd Int. Wworkshop on Patent Information Retrieval*, PaIR'09, pages 33–36, New York, NY, USA, 2009. ACM.
44. R. Rada, J. Barlow, J. Potharst, P. Zanstra, and D. Bijstra. Document ranking using an enriched thesaurus. *Journal of Documentation*, 47(3):240–253, 1991.
45. E.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the ASIS*, 27:129–146, 1976.
46. S.E. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life: Okapi at trec. *Information Processing & Management*, 36(1):95–108, january 2000.
47. T.J. Siddiqui. Intelligent techniques for effective information retrieval: a conceptual graph based approach. *SIGIR Forum*, 40(2):73–74, 2006.
48. A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. of the 19th Annual Int. Conf. on Research and Development In information Retrieval*, SIGIR'96, pages 21–29, New York, NY, USA, 1996. ACM.
49. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.A. Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel, and S. Lewis. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007.
50. J.F. Sowa. Conceptual graphs for a data base interface. *IBM Journal of Research and Development*, 20:336–357, July 1976.
51. R. Stevens, C. Goble, and S. Bechhofer. Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics.*, 1.4:398–414, 2000.
52. J. Tuominen, N. Laurenne, and E. Hyvönen. Biological names and taxonomies on the semantic web: managing the change in scientific conception. In *Proc. of the 8th Extended Semantic Web Conf. on the Semanic Web: Research and Applications - Volume Part II*, ESWC'11, pages 255–269, Berlin, Heidelberg, 2011. Springer-Verlag.
53. A. Vailaya, P. Bluvas, R. Kincaid, A. Kuchinsky, M. Creech, and A. Adler. An ar-

chitecture for biological information extraction and representation. In *Proc. of the 2004 ACM Symposium on Applied Computing*, SAC'04, pages 103–110, New York, NY, USA, 2004. ACM.

54. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.
55. E.O. Wilson. The encyclopedia of life. *Trends in Ecology & Evolution*, 18(2):77–80, 2003.
56. M. Yoder, I. Mikó, K. Seltmann, M. Bertone, and A. Deans. A gross anatomy ontology for hymenoptera. *PloS one*, 5(12), 2010.
57. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, April 2004.
58. L. Zhang and Y. Yu. Learning to generate CGs from domain specific sentences. In *Proc. of the 9th Int. Conf. on Conceptual Structures: Broadening the Base*, ICCS'01, pages 44–57, London, UK, UK, 2001. Springer-Verlag.