Proposal of Structure and Information Organization for Bullipedia

Mirjam Johannes

Master Program in
Cognitive Science and Language,
Department of General Linguistics,
Universitat de Barcelona

Under the supervision of

Dr. M. Antònia Martí Antonín
Universitat de Barcelona

Dr. Marta Vila Rigat
Unitat UB-Bullipèdia,
Universitat de Barcelona

September 2014

Contents

1. Motivation

Recently the gastronomic discourse has gained more resonance in the society. The new culinary studies such as the Bachelor's degree in Culinary and Gastronomic Sciences (interuniversity UB-UPC with CETT and Fundació Alícia) and the new projects of Ferran Adrià like Bullipedia, are only a few examples of this phenomenon. The Bullipedia project led by Ferran Adrià is promoted by elBulliFoundation at BullipediaLab, and has the academic support of the University of Barcelona and other institutions through the UB-Bullipedia Unit, situated at the Food and Nutrition Torribera Campus of the University of Barcelona.

The study presented here is dedicated to the knowledge representation in the Bullipedia encyclopaedia, an online resource that aims at containing the gastronomic knowledge of all times. The creation of the Bullipedia encyclopaedia derives from the necessity for a reliable, complete and uniform source of information in the field of gastronomy. The rise of the gastronomic and culinary discourse in the society, has increased notably the need for this kind of a resource. Bullipedia aims to provide a scientific model for all this gastronomic and culinary knowledge.

The development of the Bullipedia encyclopaedia is part of the bigger project with the same name mentioned before. The Bullipedia project has two main purposes. First, to make possible for gastronomy to become an academic discipline, in other words, to give to gastronomy an academic dimension. In order to achieve that, this project will provide reliable and structured content on gastronomy. Second, to make the content available to cooks, students, investigators, professionals and society in general through a digital tool. Although the current study is also related to the first objective, it is more strongly linked to the second one. Our purpose is to design a proposal for the Bullipedia encyclopaedia structure that would facilitate the diffusion of the Bullipedia project. The proposals of this study are of advisory nature and should be viewed as recommendations for the Bullipedia project.

Bullipedia is a multidisciplinary resource where the knowledge of various disciplines will converge and interact. Bullipedia is developed collaboratively by the culinary professionals, the academic world and up to some point the society in general. Thanks to the different profiles of its creators, the Bullipedia encyclopaedia aims at taking into account the interests of different possible users. Bullipedia will be used in university studies and in academic research related to gastronomy as well as in professional world and by society in general (Figure 1). As the contribution is mutual, Bullipedia will never be completed, but will be in continuous development.

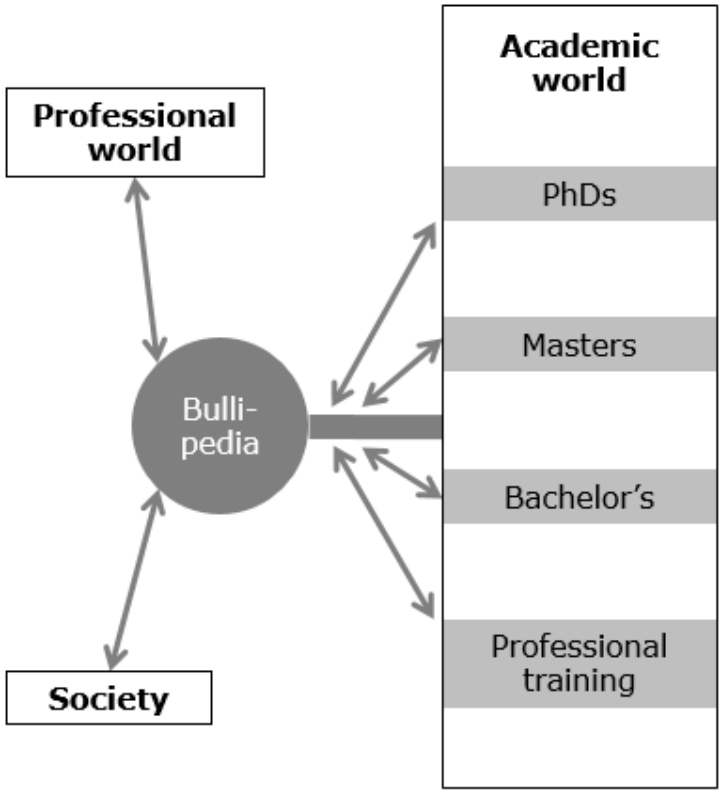


Figure 1 - Where is Bullipedia heading to? (Extracted from: *http://www.ub.edu/campusalimentacio/ub-bullipedia/ca/capon.html*)

It should be also pointed out that the academic world of Bullipedia counts with a wide variety of disciplines from art to chemistry, documentation, terminology, history, etc. Therefore, the interdisciplinarity and multifaceted nature of Bullipedia are omnipresent in its creation process.

The difficulties this research has to face are related to Bullipedia's special nature:

- **Bullipedia will be a multifaceted resource.** Saying that Bullipedia will include knowledge of various disciplines all related to the world of gastronomy, leaves no doubt that we are dealing with a multifaceted resource. The wide range of information that goes from the biology perspective on non-elaborated products to the artistic perspective on certain final elaborations, passing through chemistry, history and other disciplines related to food, to set out only some examples, requests the resource to be comprehensive but specific enough to deal accurately with the culinary world and all the other related fields.
- **Bullipedia will have wide variety of final users and purposes.** Our proposal has to be general enough to be able to integrate different kinds of knowledge and to function as a resource for different groups of users. Therefore, our proposal has to be designed to help people during the culinary creation process, it also has to be functional as a tool for gastronomic studies etc.

So, on the one hand, Bullipedia has to be specific enough to cover all the topics, and on the other hand, it has to be general enough to be functional for various users and purposes. We will keep these requisites in mind when developing the model of knowledge representation.

Our study is supported by different lines of research. The theory of knowledge representation and reasoning that has grown out of the theories of artificial intelligence gives us the criteria to construct a knowledge base. Semantic networks will serve as a model of knowledge representation, because we believe that treating the knowledge present in Bullipedia as a set of concepts that are related to one another will facilitate the understanding of the culinary world. The theoretical basis of our study is laid by lexical semantics. We will also treat two cases as direct references: WordNet as an example of lexical knowledge base and Wikipedia as an example of the internal structure of a digital encyclopaedia. As mentioned above, the current research deals only with the sketch of a knowledge representation model and does not contribute to the encyclopaedic content.

This research is organised as follows: in Section 2 we will present basic concepts of two disciplines, knowledge representation and lexical semantics that are at the basis of any kind of knowledge representation system, and also we will make a critical presentation of two reference models: WordNet and Wikipedia. In Section 3 our proposal is presented: we will enter into the world of Bullipedia and introduce its unique properties. In Section 4 we will arrive to the conclusion and discuss the future research.

2. Basic concepts

We can define Bullipedia as a culinary knowledge base. And as Bullipedia will be the first one of this kind and there is no equivalent point of reference, we will follow the lead of three lines of research related to our purpose. In what follows, we present the basic concepts in these areas that will help us to formalize our proposal.

- First, in Section 2.1. we introduce the area of knowledge representation in artificial intelligence and take a look at the functioning of semantic networks. The theory of knowledge representation forms the scientific basis of this research.
- Second, in Section 2.2. we examine the linguistic foundations of lexical semantics and treat WordNet as one of its most remarkable representatives. As the theory of lexical semantics has been the theoretical basis of knowledge representation in artificial intelligence, we also consider it as important theoretical knowledge for our proposal.
- Finally, in Section 2.3. we present a critical analysis of Wikipedia, the online encyclopaedia that is also one of the points of reference and inspiration in building Bullipedia. We will discuss the pros and cons of the Wikipedia's model in order to define our proposal. As in the case of Wikipedia, in the case of Bullipedia we also have to deal with multifaceted entities where objects may be represented from diverse, distinct ontological perspectives with each perspective describing different states of an object within the same application domain.

2.1. Knowledge representation[1]

Knowledge representation and reasoning (KR) is the field of artificial intelligence (AI) devoted to representing information about the world in a form that a computer system can utilize it to solve complex tasks. KR is the study of thinking as a computational process. Knowledge representation and reasoning have been placed at the centre stage in AI research. The current AI theory and practice dictate that intelligent systems have to be knowledge based, and therefore, the suggestions that AI can be called applied epistemology should be taken into account. Nevertheless, defining the task and methods of knowledge representation still stirs considerable controversy and there is little agreement on how the problem has to be solved.

In short, AI systems are composed by a knowledge base with facts about the world (knowledge representation) and rules and an inference engine that applies the rules to the knowledge base in order to answer questions and solve problems (reasoning). In our research the part of automatic reasoning is substituted by an interface and a query system, as the users of Bullipedia take the place

[1] This chapter is based on Shapiro (1991)

of the reasoning machine. Therefore, our interest for KR is more related to the construction of a knowledge base.

The knowledge base is a symbol structure representing a collection of facts about a domain of discourse. At the very least, it would be expected that a knowledge representation system have to offer facilities for constructing and querying the knowledge base. In our case that means offering to the users a clear representational screen and an easy way to consult it.

Proposed notations of knowledge representation are classified into three basic paradigms: logic-based, procedural, and semantic network. In order to meet the aims of our research and introduce our final proposal, we will concentrate on the semantic networks as the most wide-spread model for knowledge representation.

2.1.1 Semantic Networks

Some authors proclaim that semantic networks simply offer a graphical notation for logical formulas, while others argue that semantic networks offer a fundamentally different representational paradigm. Organizing the knowledge included in a knowledge base is important from a cognitive science viewpoint because human memory also seems to be highly structured. Thus, it can be said that semantic networks were originally motivated by cognitive models of human memory.

In semantic network, knowledge is usually represented on a labelled, directed graph whose nodes represent concepts and entities in the domain of discourse, while its arcs represent relationships between these entities and concepts. In the case of the semantic networks the most important structuring mechanisms or relations that have been adopted for knowledge organization are:

- *instance-of* (or classification)
- *is-a* (or generalization)
- *part-of* or *member-of* (or aggregation)

In Figure 2 we can see the oldest known semantic network: Porphyry's tree that was drawn in the 3rd century AD by the Greek philosopher Porphyry in his commentary on Aristotle's categories. Despite its age, the tree of Porphyry represents the common core of all modern hierarchies that are used for defining concept types. In Figure 2 we can see how all the classes and their subclasses are in *is-a* relation, meanwhile the *instance-of* applies only to individuals belonging to a class (like Plato, Socrates, etc.) and does not admit subclasses.

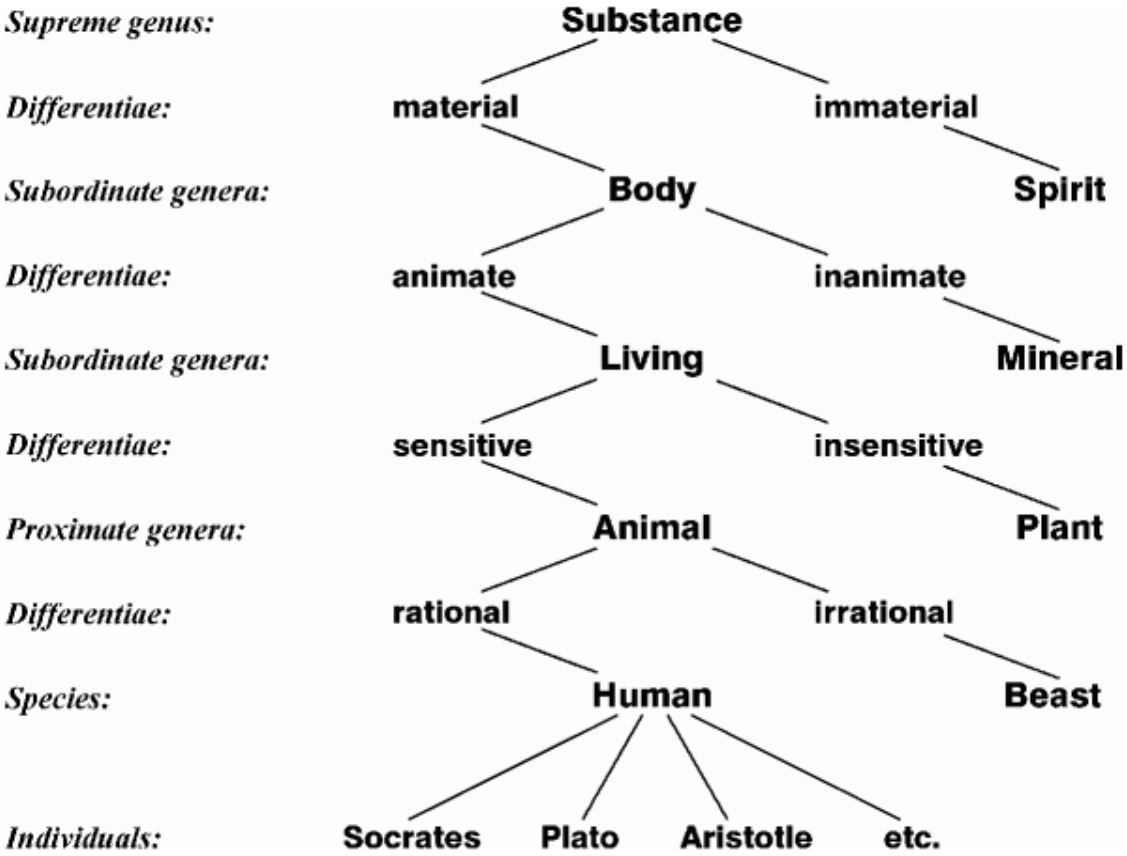


Figure 2 - Tree of Porphyry (Shapiro 1991: 1494)

Is-a and *instance-of* represent inheritance and are popular in many knowledge representation schemes. Inheritance is crucial in AI because it means systems can exploit taxonomic reasoning. With inheritance, the information is distributed through different levels, avoiding redundancy (Malrieu 2002:161).

Some problems that have risen in research of AI that are particular to knowledge representation systems are:

1. **Inconsistency** – a knowledge base should be considered inconsistent if it is possible to derive contradictory conclusions from it. Many times inconsistency is the result of a mismatch between old knowledge already existing in the knowledge base and the new inserted knowledge.
2. **Incompleteness** – a knowledge base might have incomplete information about their universe of discourse. Incompleteness might arise due to insufficient information about that universe. In fact, all KB are in some way incomplete.

Incompleteness and inconsistency pose serious problems and should be avoided in every sense in the development of knowledge bases.

With respect to our proposal for Bullipedia, we aim to learn our lessons from previous researches, in order to avoid the above-mentioned problems. Our purpose is to come up with an efficient model of knowledge representation that meets all the requisites of Bullipedia. Hence we should take into account the following suggestions:

a. The categorization and inheritance hierarchy are important organizers of every knowledge base. In Bullipedia, every entity placed at any level of the hierarchy should be easily accessible in its categorization tree in order to represent clearly the inheritance of its characteristics.
b. Inconsistency should be avoided by providing a fixed template and general organization model. Using a uniform format helps to avoid inconsistency and incompleteness of the resource.

In conclusion it can be said that even though many of the controversies about what knowledge representation is or does remain, a number of features and proposals within this domain should be considered when building Bullipedia.

2.2. Lexical semantics

Lexical semantics is a subfield of linguistic semantics. It is the study of what individual lexical items mean, why they mean what they do, how we can represent them, and where the combined interpretation for an utterance comes from. We are especially interested in lexical semantics because it is frequently used as the basis for the structure of knowledge bases and has played a fundamental role in the development of knowledge representation ontologies, dictionaries, etc. Lexical semantics assumes that a word (concept) is a conventional association between lexicalized concept (meaning) and an utterance (form). In other words, the starting point for lexical semantics is the mapping between word forms and word meanings. One question that lexical semantics explores is whether the meaning of a lexical unit is established by looking at its neighbourhood in the semantic net, by looking at the other words it occurs with in natural sentences, or if the meaning is already locally contained in the lexical unit. It is important to keep in mind that a person does not experience the word form and meaning as two separate things, but as two aspects of a unitary phenomenological entity.

2.2.1. WordNet[2]

[2] This chapter is based on Miller & Fellbaum (1991) and Beckwith et al. (1991).

WordNet is a lexical database for English and many other languages that we should take into consideration when constructing Bullipedia. This lexical resource is based on psychological principles in that it instantiates a structure postulated for the mental lexicon based on the results of psycholinguistic research. Therefore, it is a psychologically motivated model. This representation offers an opportunity to clearly distinguish a system of semantic relations between forms and the system of semantic relations between meanings. For example some forms have several different meanings (polysemy and homonymy), and some meanings can be expressed by several different forms (synonymy).

WordNet is organized by semantic relations. The relations that are especially significant for the structure of the lexicon are relatively small in number, therefore, the database of WordNet is restricted to the relations like synonymy, antonymy, hyponymy, hypernymy, and three types of meronymy and holonymy. The basic unit of WordNet is the synset, a list of (quasi) synonyms that represent a concept. Every synset has a gloss that defines its content. Synsets are connected to other synsets by pointers representing relations such as the ones mentioned above. Therefore, the main relation among words in WordNet is synonymy, as each entry is a unique synset connected to other synsets. Words in one synset denote the same concept and are interchangeable in many contexts.

In what follows, we are going to list the most frequently encoded relation among synsets in WordNet. The majority of the WordNet's relations connect words from the same part of speech (POS). Thus, WordNet really consists of four sub-nets, one for nouns, one for verbs, one for adjectives and one for adverbs, with cross-POS pointers. Each POS has its own specific kind of relationships.

a. Polysemy (and homonymy)– is a relation where one sign can have many possible meanings. WordNet represents polysemy distinguishing each meaning by means of different synsets. As we can see in Figure 3 in the case of *asparagus* the network gives us two possible meanings: first, *asparagus* as plant and, secondly, as the edible part of the plant. The relation of polysemy is common in case of all four sub-nets.

Noun

- S: (n) **asparagus**, edible asparagus, Asparagus officinales (plant whose succulent young shoots are cooked and eaten as a vegetable)
- S: (n) **asparagus** (edible young shoots of the asparagus plant)

Figure 3 - Synsets of *asparagus*

b. Hyponymy/hypernymy - in the case of nouns, hyponymy and hypernymy are the most common relations. These relations create a hierarchical tree structure, i.e., a taxonomy. A hyponym anywhere in the hierarchy can be said to be *a kind of* all of its superordiantes. The direct hypernym of *asparagus* (plant) is *herb* and through inherited hypernymy it is related to all the hierarchy of hypernyms of *herb*: *vascular plant*, *plant*, *organism*, *living thing* etc. (Figure 4).

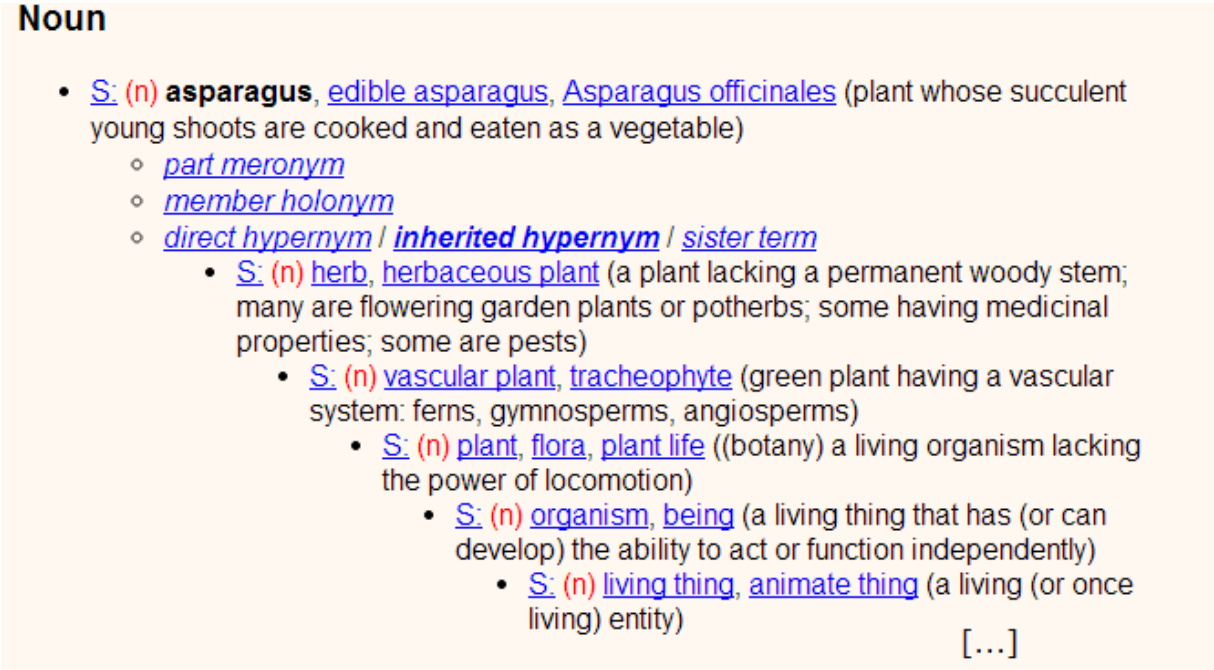


Figure 4 – Inherited hypernymy of *asparagus* (plant)

c. Meronymy/holonymy – mark the relationship between a term denoting the whole and a term denoting a part of it. As we can see in Figure 5, WordNet distinguishes between two kinds of meronymy/holonymy: *part meronym* makes reference to the part-whole relation, example of which are a plant and its edible parts; *member holonym* makes reference to a family where the asparagus as a plant belongs. It marks that the plant of asparagus is a *member-of* genus Asparagus. [3]

[3] Usually the difference between these two types of meronymy/holonymy is explained by the examples of *wheel – bicycle* (where *wheel* is "part-of" bicycle) and *violinist – orchestra* (where *violinist* is a "member-of" *orchestra*).

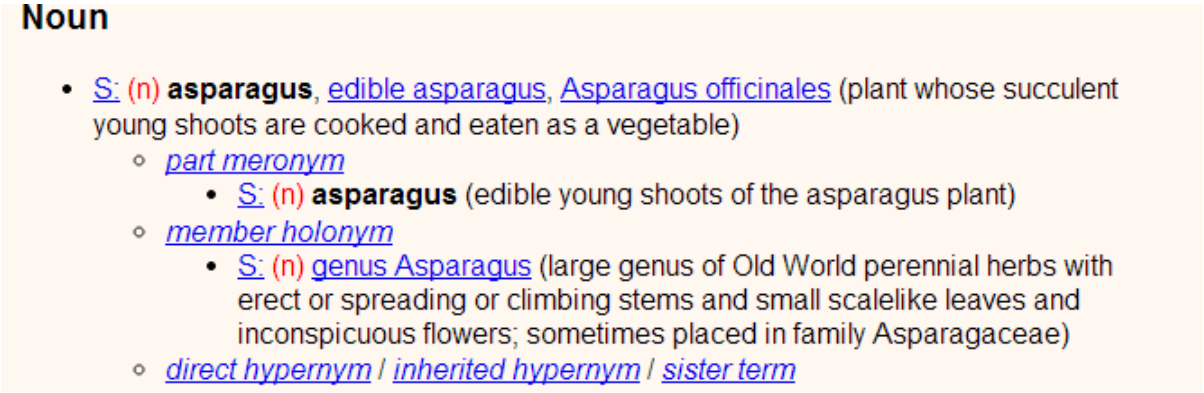


Figure 5 – Meronymy/ holonymy relations of *asparagus* (plant)

d. Sister term – is another interesting relation represented in WordNet. This relation is used to designate a pair of synsets that share a hypernym. We can see some examples of the sister terms of *asparagus* (plant) in Figure 6. They are entities that are also described as *kind-of herb,* like for instance, *barrenwort, mayapple* etc. As *asparagus* (edible part) has different hypernym than *asparagus* (plant) they have also different sister terms as we can see in Figure 6.

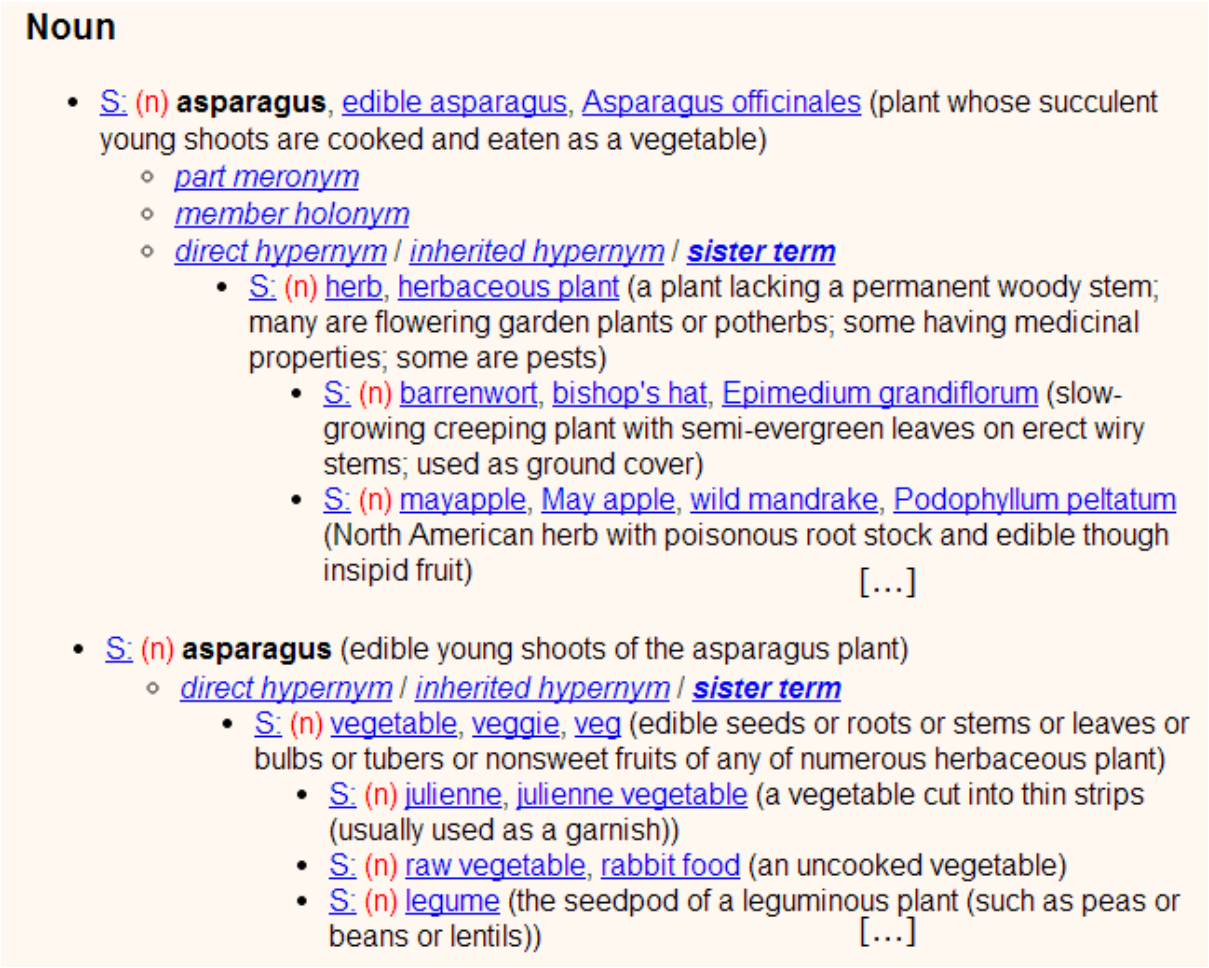


Figure 6 – Sister terms of *Asparagus* (plant) and *Asparagus* (edible part)

e. Antonymy – is rarely represented among the nouns and is mostly used for representing relations between adjectives. For example, in case of *boiled* WordNet gives us the antonym *raw* (Figure 7).

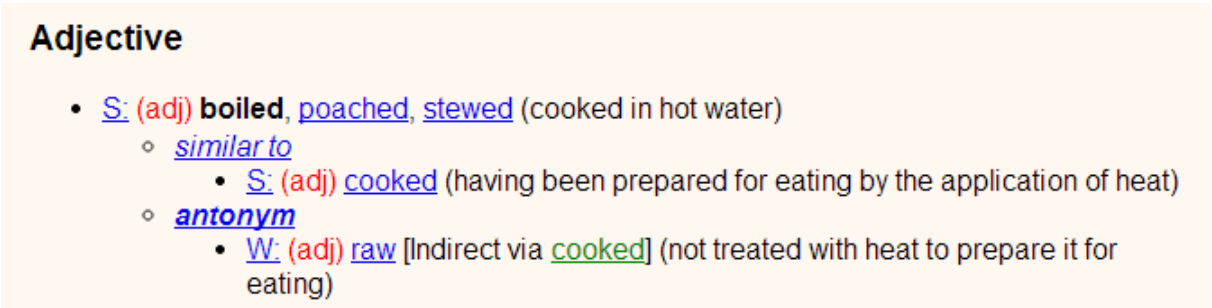


Figure 7 – Adjective *boiled*

f. Troponymy – is a specific relation of verbal entries in WordNet. It was proposed by Miller and Fellbaum (1991) for WordNet, in order to describe the relation of "manner-of" between two lexemes. "A troponym, then, is a verb that constitutes an

elaboration of its superordinate by expressing a particular way or manner in which the activity referred to by the superordinate verb is carried out" (Beckwith et al. 1991: 221). As we can see in Figure 8 the troponyms of the entry *fry* are *stir fry*, *sauté*, *deep-fry* etc. and they represent the different ways to *fry.*

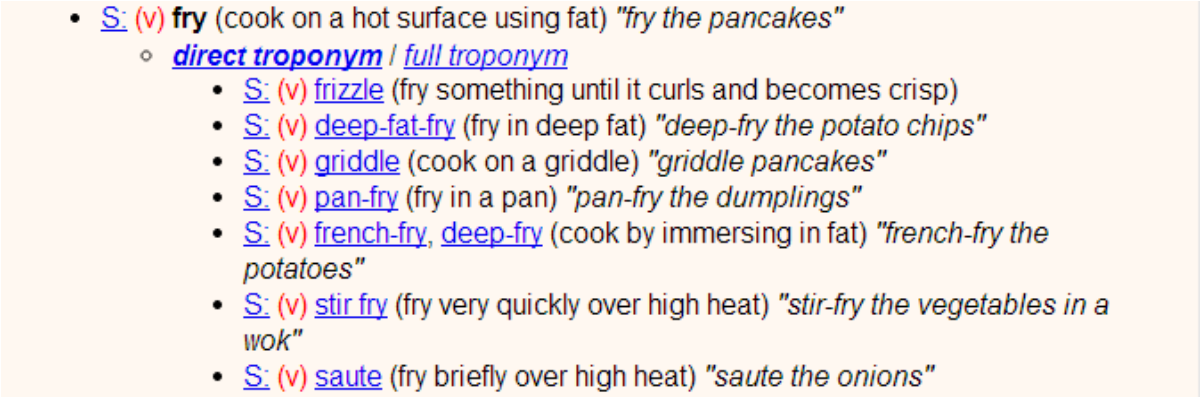


FIgure 8 – Troponymy relation of verb *fry*

WordNet could be a complementary source of information to Bullipedia. Given that the WordNet's inner structure is based on semantic relations, it could give a different point of view to Bullipedia which will count with more a encyclopaedic structure.

Moreover, the meaning and form mappings established in WordNet could be used as a possible model for the structure of Bullipedia for many reasons:

a. It is a system that has been developed since 1985 and reflects therefore the outcomes of many years of research.
b. WordNet is based on the knowledge acquired over the years about how human beings process language and store knowledge about language and is therefore a justified proposal of knowledge representation.
c. The hierarchical organization of nouns based on *a kind of* (hyponymy/hypernymy) and *part of* (meronymy/holonymy) relations is also common among the entities of Bullipedia. The world of Wikipedia[4]

Wikipedia is a collaboratively edited, multilingual, free Internet encyclopaedia. Currently it is the 6th most popular website and the most widely used encyclopaedia in the world (Lehmann 2012: 2). Wikipedia is not just an encyclopaedia but can be viewed as anything from a corpus, taxonomy, thesaurus, hierarchy of knowledge topics to an ontology. Wikipedia has revolutionized our point of view on the nature of knowledge from something that was purely academic to something that could

[4] This chapter is based on Medelyan et al.(2008)

be described as a general belief of the community. Moreover, Wikipedia proposes its own account of truth, where the beliefs are to be understood as knowledge due to their usefulness, public character and future developments. For our research Wikipedia is particularly important as a possible model of knowledge representation.

Wikipedia has adapted some of the features of paper encyclopaedias for online environment as well as it has its own unique features arisen during the Wiki editing process. In section 2.3.1, we present the most emblematic features of Wikipedia. In 2.3.2, we discuss the main drawbacks present in the Wikipedia model. The drawbacks in 2.3.2. are directly related to the features of Wikipedia presented before.

2.2.2. Features of the Wikipedia structure

(I) Article

It is the basic unit of information in Wikipedia. Each article describes a single concept, or in other words, there is a single article for each concept. The articles begin with a brief overview of the topic and the first sentence defines the entity and its type. As we can see in Figure 9, *asparagus* is defined as "a spring vegetable, a flowering perennial plant species in the genus *Asparagus*", where we can identify its type (vegetable, plant species). In order to maintain the uniformity and have single article for one concept, Wikipedia guideline recommends to use redirects to link equivalent terms to the preferred article title. For example, regardless of whether we are looking for *asparagus* or *asparagus officinalis* in the Wikipedia search engine, it will direct to the same article page.

Asparagus

From Wikipedia, the free encyclopedia

For the botanical genus, see Asparagus (genus). For the colour, see Asparagus (color).

Asparagus officinalis is a spring vegetable, a flowering perennial[1] plant species in the genus *Asparagus*. It was once classified in the lily family, like its *Allium* cousins, onions and garlic, but the Liliaceae have been split and the onion-like plants are now in the family Amaryllidaceae and asparagus in the Asparagaceae. *Asparagus officinalis* is native to most of Europe, northern Africa and western Asia,[2][3][4] and is widely cultivated as a vegetable crop.

Figure 9 - *Asparagus* disambiguation and brief description

(II) Disambiguation page

According to Wikipedia Manual of Style[5] disambiguation page is designed to help a reader find Wikipedia articles on different topics that could be referenced by the same search term. Disambiguation pages are Wikipedia's tool to deal with polysemy. In some cases, the search takes user directly to the disambiguation page, in other cases, the search redirects the user directly to the

[5] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Disambiguation_pages

most searched article. In the latter case, the first line of the article links to the disambiguation page or links directly to other same-named concepts. For example, from the article *asparagus* (Figure 9), we can access directly to "*Asparagus (genus)"* that stands for the botanical family and "*Asparagus (colour)"* that stands for the particular green colour of the plant.

(III) Category structure

The central goal of the category system is to provide navigational links to all Wikipedia pages in a hierarchy of categories, which readers can browse and quickly find sets of pages on topics that are defined by those characteristics. In short, categories are nodes for organizing the articles they contain, but usually categories themselves are not articles. Authors of Wikipedia are encouraged to assign categories to their articles. The goal of the category structure is to represent the hierarchically organized information. As we can see in Figure 10 that depicts the category structure of the article *Asparagus,* categories are situated at the bottom of the page and, in this case, include the notions *Asparagus, Medical plants of Africa, Medical plants of Asia, Medical plants of Europe, Stem vegetables, Perennial vegetables, Plants with indehiscent fruit* and *Flora of Nepal*.

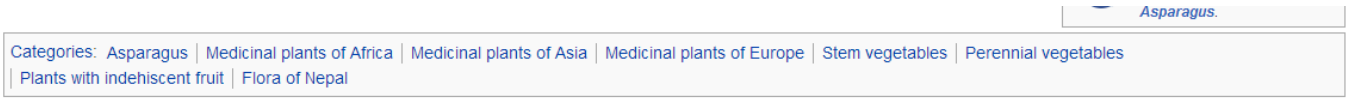


Figure 10 – *Asparagus* categories

(IV) Infobox

Wikipedia's infobox is a type of template that displays factual information in a structured uniform format. Its objective is to summarize the key facts that appear in the article. According to Medelyan (2008) the generalized infobox feature grew out of the original taxoboxes (taxonomy infoboxes) that editors developed to visually express the scientific classification of organisms. Adding an infobox to articles facilitates the retrieval of most important characteristics and facts of the entity.

In Figure 11 we can see two examples of infobox: *brie* and *pecorino.* The most important information we can retrieve from these infoboxes is about the country of origin, source of milk and texture of the cheese.

Figure 11 – Infoboxes of *Brie* and *Pecorino*

After this short list of Wikipedia's main features, we will move to the drawbacks that we have detected.

2.2.3. Drawbacks of the Wikipedia structure

Even though Wikipedia structure is well-defined in user manuals, the irregular practice in applying user guidelines causes incompleteness and inconsistency. There are many detectable deficiencies in the actual representation of many previously mentioned features. In the following part we are going to discuss those problems in more detail:

(I) Articles

Not all articles deal with only one single concept. As in some cases the difference between polysemous entities is not well defined, one article may deal with two different concepts. For instance, there is a mismatch between the title of the article (*Asparagus*) and the entity (*Asparagus officinalis*) defined in the first sentence (Figure 9).

The definition of the concept given in the first sentence has no uniform practice and is often not informative enough. Given that the first lines of an article count as a brief description that situates the users within the scope of the article, we should be critical to the information represented in this paragraph. In Figure 12 we have examples of definitions given to two different concepts: *asparagus*

and *agaricus bisporus.* Comparing these two definitions it becomes obvious that even though the entities resemble in many aspects and share the same properties, they are defined through different aspects. For example, both of them mention where the plant is native to, but in case of *asparagus* it is said in the end of the definition and in case of *agaricus bisporus* it is said in the first phrase. In definition of *asparagus* there is more importance placed on how the plant used to be classified, that in our opinion is less relevant.

Asparagus officinalis is a spring vegetable, a flowering perennial[1] plant species in the genus *Asparagus*. It was once classified in the lily family, like its *Allium* cousins, onions and garlic, but the Liliaceae have been split and the onion-like plants are now in the family Amaryllidaceae and asparagus in the Asparagaceae. *Asparagus officinalis* is native to most of Europe, northern Africa and western Asia,[2][3][4] and is widely cultivated as a vegetable crop.

Definition Asparagus

Agaricus bisporus is an edible basidiomycete mushroom native to grasslands in Europe and North America. It has two color states while immature—white and brown—both of which have various names. When mature, the same mushroom has yet another popular name.

When immature and *white*—this mushroom may be known as **common mushroom**, **button mushroom**, **white mushroom**, **cultivated mushroom**, **table mushroom**, and **champignon mushroom**. When immature and *brown*—this mushroom may be known variously as **Swiss brown mushroom**, **Roman brown mushroom**, **Italian brown**, **Italian mushroom**, **cremini** or **crimini mushroom**, **brown cap mushroom**, or **chestnut mushroom**.[2]

When mature, the same mushroom is known as **Portobello mushroom**.

Agaricus bisporus is cultivated in more than seventy countries,[3] and it is one of the most commonly and widely consumed mushrooms in the world.

Definition Agaricus bisporus

Figure 12 -*Asparagus* and *Agaricus bisporus* definitions

(II) Disambiguation pages

In Wikipedia the polysemy is not always resolved by disambiguation pages and therefore some cases of inconsistency may appear. In the case of *asparagus* where we have only three meanings to disambiguate the links are preferred (Figure 9), but in case of *honey* where there are dozens of different meanings, the disambiguation page is preferred. Probably, in many cases this irregularity is due to the different number of polysemous cases. One of the two possibilities should be preferred and given that in case of many possible meanings it is complicated to fit them in the top of the article, a separate page would be more adequate.

(III) Category Structure

Wikipedia's category system does not offer a complete and adequate overview of the hierarchy of entities. The subjects of Wikipedia accept multiple ways of being categorized. Consequently, in many cases, the categories assigned to an entity are far from being adequate. In the categorization guidelines[6] Wikipedia states that "the central goal of the category system is to provide navigational links to all Wikipedia pages in a hierarchy of categories which readers, knowing essential

[6] http://en.wikipedia.org/wiki/Wikipedia:Categorization

characteristics of a topic, can browse and quickly find sets of pages on topics that are defined by those characteristics." In other words, they are used to link concepts with similar ones, with sister terms that share the same hypernym or category. The fact that, for example *Asparagus* (Figure 10) is tagged in three different categories for medical plants *(Medical plants of Africa*, *Medical plants of Asia*, *Medical plants of Europe*) is very misleading, given that inside the article the possible medical uses of the plant are mentioned only once. Also, the category of *Flora of Nepal* is confusing because it might give an impression that asparagus only grows in Nepal, as it is not included to any other category that would define it as a plant of some other country.

Wikipedia's category system is currently incapable of supporting the search for entities that share more than one category. At present it offers only free-text search capabilities for the users and therefore it is difficult to find, for example, all the stem vegetables that are also medical plants. That is because, not all plants that meet the criteria are tagged in both categories and there is no such category as "medical stem vegetables". Therefore it could be said that the current category system is inconsistent and is not contributing to the easy access to the information.

(IV) Infoboxes

Currently, the **Wikipedia articles consist mostly of free text and there are some types of structured data such as infoboxes, tables, lists.** The more structured the information is, the easier it is to achieve its uniformity and make it automatically accessible. Therefore, one of the drawbacks of Wikipedia is the lack of structured data. There are projects that extract information from Wikipedia in order to store it in formats accessible to database applications. For example, DBpedia (Lehmann 2012) is a project that extracts structured data from Wikipedia and turns it into a rich knowledge base. Its aim is to create an entirely new ontology by harvesting facts from Wikipedia. The facts are stored as a vast set of RDF triples. Each article in Wikipedia, or more precisely each infobox, becomes an entity in DBpedia.

There are cases of inconsistency between infoboxes. Wikipedia guidelines[7] on infoboxes list different cases of inconsistency that may occur: historical incompleteness, hierarchical inconsistency, feature inconsistency and lack of information. In most cases the inconsistency factors of an infobox are the same that those of an article. For example, historical incompleteness occurs because certain desired information may simply have been lost over time and therefore the knowledge we have of medieval dishes is much poorer than we have of modern cuisine. Feature inconsistency occurs when some optional features are not listed in an infobox. In Figure 9 we have an example of two products

[7] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Infoboxes

that share the same infobox, however, in the infobox of *pecorino* many of the features that appear in the infobox of *brie* are not listed. As the feature inconsistency is more a matter of the lack of uniform practice it should be easier to avoid it than historical incompleteness. When filling in the fields of an infobox, all fields should always be completed. Nevertheless, the inconsistency could be avoided already beforehand valuating the importance of each optional field at the moment of the creation of an infobox. It should also be noted that even though infobox is one of the emblematic features of Wikipedia not all articles have infoboxes.

Different attribute names are often used for the same kind of entries and different names are given to same type of values. For example, there are many different attribute names to represent the date and location of birth and death. Also, the same values that correspond to an attribute might have different names (Vila et al. 2013). We can see in Figure 13 that to name the course that comes before the main course three different terms have been used: *antipasto*, *appetiser* and *hors d'oeuvre*. This inconsistency in naming can be probably explained by the fact that these three dishes are all remarkable representatives of three different national cuisines. And in Italian, Spanish and French cooking traditions the course before the main course have different names. However, in order to avoid confusion one of the names should be preferred or a more general one should be used.

Figure 13 – inconsistency of values

To sum up what we have learned about Wikipedia, it could be said that even though there are many inconsistencies in the current version of this popular resource, we find the unique features of Wikipedia really useful and functional for any other online encyclopaedia.

In contrast to previous sections on KR and lexical semantics, the present one does not end with recommendations for Bullipedia. The recommendations will be given in the next section as part of our proposal. This division can be explained by the special link between Bullipedia and Wikipedia. After all, Wikipedia is one of the main points of reference in development of Bullipedia.

3. The world of Bullipedia.

The world of Bullipedia is complex and unique. Its approach to gastronomic knowledge and ambition to give gastronomy an academic dimension make of Bullipedia a resource with special requirements. It has been said above that Wikipedia revolutionized our point of view on the nature of knowledge from something that was purely academic to something that could be described as a general belief of the community. But Bullipedia will go beyond that, it aims converge the scientific knowledge with the knowledge of professionals of the field and take also into account the general beliefs of the society. In this chapter we are going to discuss Bullipedia in more detail. In 3.1. we discuss the nature of the Bullipedia encyclopaedia and the Bullipedia project in general. In 3.2. we present our proposal for the Bullipedia encyclopaedia.

3.1. Introduction to Bullipedia

What is Bullipedia?

When the creation of the Bullipedia encyclopaedia was announced to public, it was defined as "an online database which [would] hold every bit of gastronomic knowledge ever uncovered"[8]. In this sense, Bullipedia was defined as a professional tool on cooking and gastronomy. Nevertheless, Ferran Adrià and his team soon realized that they had to go a step further and that focusing on the online tool was not enough. Now, creation of the Bullipedia encyclopaedia forms part of a bigger project with the same name[9]. Ferran Adrià (2014: 15) listed the objectives of the Bullipedia project as follows:

- to organize all culinary knowledge in a clear, orderly and concise form
- to create an internet tool that allows access, sorting, use, and exchange of all this knowledge. The idea was inspired by search engines as Google and Yahoo and Internet encyclopaedias – Wikipedia
- to propose a model that could also be used by other disciplines.

Also, the Bullipedia project is strongly linked to the academic world. In this context, it aims to give gastronomy an academic dimension within universities. In order to achieve this goal, Bullipedia has to provide a complete, structured and validated gastronomic content.

What are its possible uses?

[8] http://multisite-blog.digital.telefonica.com.s3.amazonaws.com/wp-content/uploads/2012/10/Bullipedia-Factsheet.pdf

[9] In the following chapter the name *Bullipedia* will make reference to the encyclopaedia.

As explained at the beginning of this work, Bullipedia is a multifaceted resource that will contain knowledge from multiple disciplines. The possibility to access to a wide variety of information increases the group of its possible users as well as guarantees many potential uses. We can observe in Figure 1 how Bullipedia will be used in university studies as well as by culinary professionals and society in general. The contribution of these three pillars towards Bullipedia is intended to be simultaneous and mutual. As the contribution is mutual, Bullipedia will never be completed, but will always continue developing.

What knowledge does the Bullipedia project consists of?

The Bullipedia project is composed by ten different lines of research. In each line of research, experts from different disciplines work together, making Bullipedia an interdisciplinary project by nature. We will shortly introduce these ten lines of research represented in Figure 14:

- **Creative Process** aims to map the creative processes behind different disciplines in order to help us understand the elements that intervene and to have a better overview of the creative activity.
- **Art and Cooking** aims to study intersections and disruptions between cooking and art. It also examines how this question has been viewed in philosophy.
- **Tools** aims to analyse tools used in cooking.
- **History** aims to collect and organize the history knowledge related to gastronomy.
- **Products** aims to create a classification of culinary products. At the same time it also analyses the culinary products from scientific and gastronomic points of view.
- **Documentation** aims to collect and organize knowledge and documentation about food and, done that, to create a platform for the diffusion of gastronomic knowledge.
- **Organization and management** applies general management knowledge to the gastronomy field.
- **Science and cooking** aims to do research in the field where science and cooking meet.[10]
- **Technics and Technology** analyses the techniques used in cooking

[10] The course on science and cooking in Harvard University is especially relevant in this sense (http://www.seas.harvard.edu/cooking).

- **Terminlogy** analyses the new terms and new senses that have arisen in the framework of this project

Figure 14 - Lines of research (Extracted from: http://www.ub.edu/campusalimentacio/ub-bullipedia/ca/projectes.html)

3.2. Our proposal for Bullipedia

The main questions we are dealing with in this section are:

a. How should we distribute the information in Bullipedia articles?

b. In which format should the information be represented? Our list of formats is the result of the enrichment of the proposal by Català (2013: 119-120). The overview of possible eight formats is given here:

- Free text without any restrictions of length
- Predefined textual information where some attributes have predefined values from a list. One or more than one items from a list can be selected.
- Numerals
- Boolean system
- Dates in an uniform format
- Graph, an ontology where a user can navigate and obtain more information during the process.
- Image
- Attribute-value is a relationship where the properties are defined by a pair of attribute and value. The value can be in any of the previous formats.

In order to answer the previous questions we take into consideration the drawbacks set out in the previous sections. Our proposal for the organization and structure of Bullipedia is rooted in the research done in the fields of knowledge representation, lexical semantics and on Wikipedia. In what follows we summarize some recommendations derived from this background for the Bullipedia's entry:

1. **Simple name in the title and in the article**

   Each Bullipedia entity begins with a title (Figure 15). The use of simple names for the entry whenever possible will guarantee the consistency in naming. The use of scientific names and less common popular names should avoided. Therefore, as we can see in Figure 16 a simple name is used in the title and in the definition.

2. **Access to the polysemous terms via disambiguation page**

   Given that the presence of polysemy in the Bullipedia is inevitable, we have to find the best way to deal with it. As concluded in the 1.3.2 (II) a separate disambiguation page would be the most universal solution. The link to the disambiguation page should be placed at the right of the title (Figure 15 and 16). However, when the entity does not have any polysemous interpretation, there is no link to the disambiguation page (Figure 16).

3. **Basic dictionary definition of the entity**

   Each concept should be followed by plain dictionary meaning in the format of free text (Figure 15 and 16).

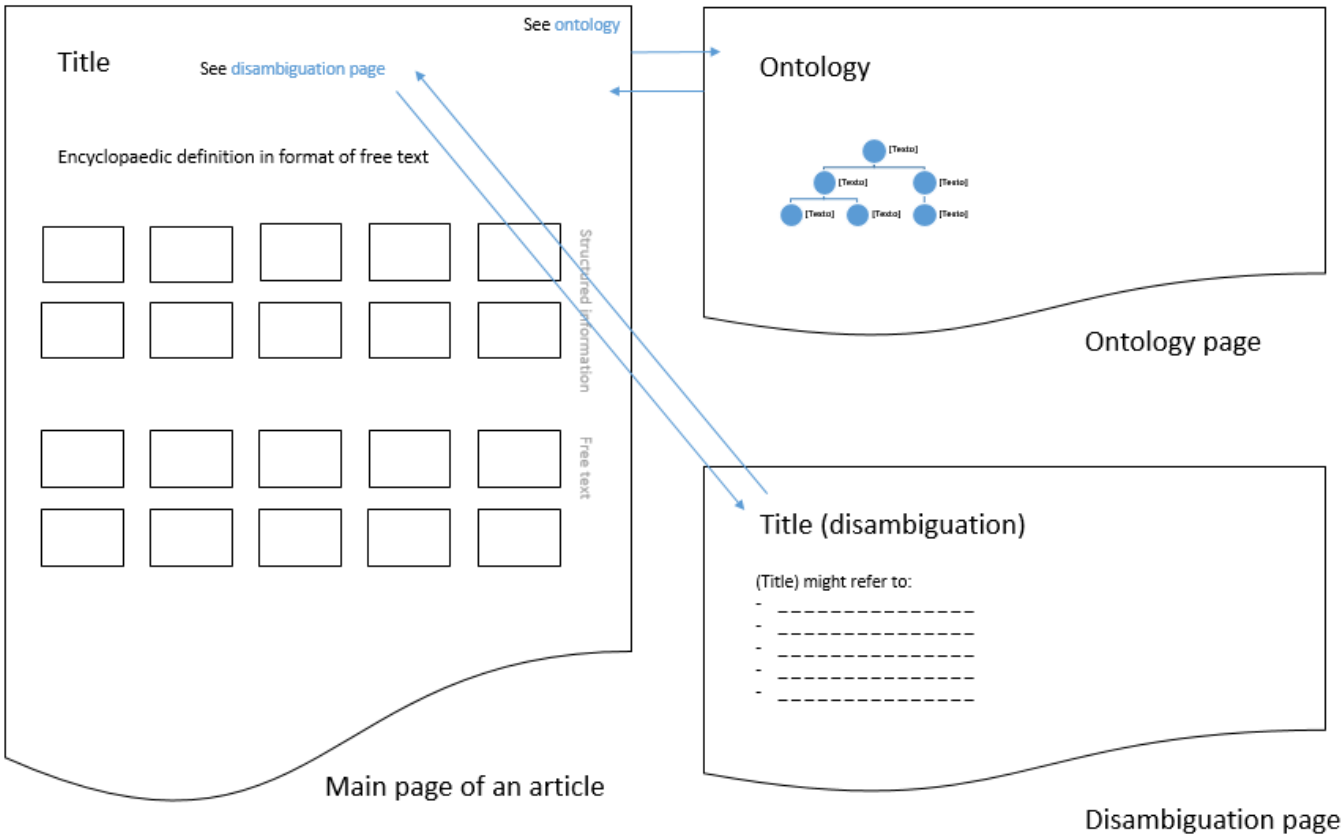


Figure 15 - Bullipedia general proposal

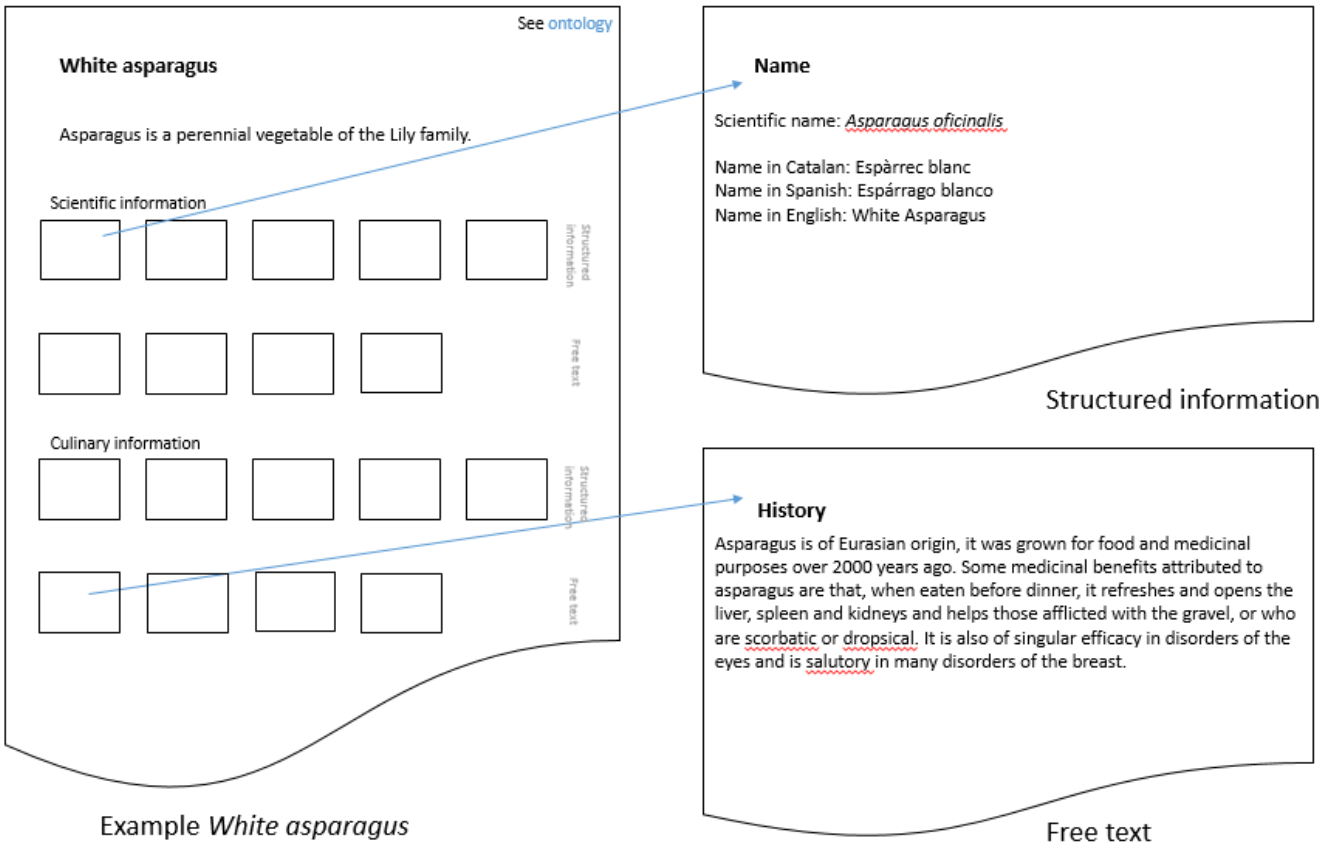


Figure 16 - Example *Esparrago blanco mediano grande*

4. **Link to the ontology where we can see the node in the hierarchy of categories the entity belongs.**

   As result of the drawbacks of Wikipedia categories (1.3.2. (III)), we propose that the category hierarchy should be present in each entry in order to provide easy access to new knowledge and localize the entity with respect to the other entities. Within the Bullipedia project, a classification of the gastronomic knowledge is being developed (left side of Figure 17).[11] Also, a classification of culinary products is being created (right side of Figure 17).[12]

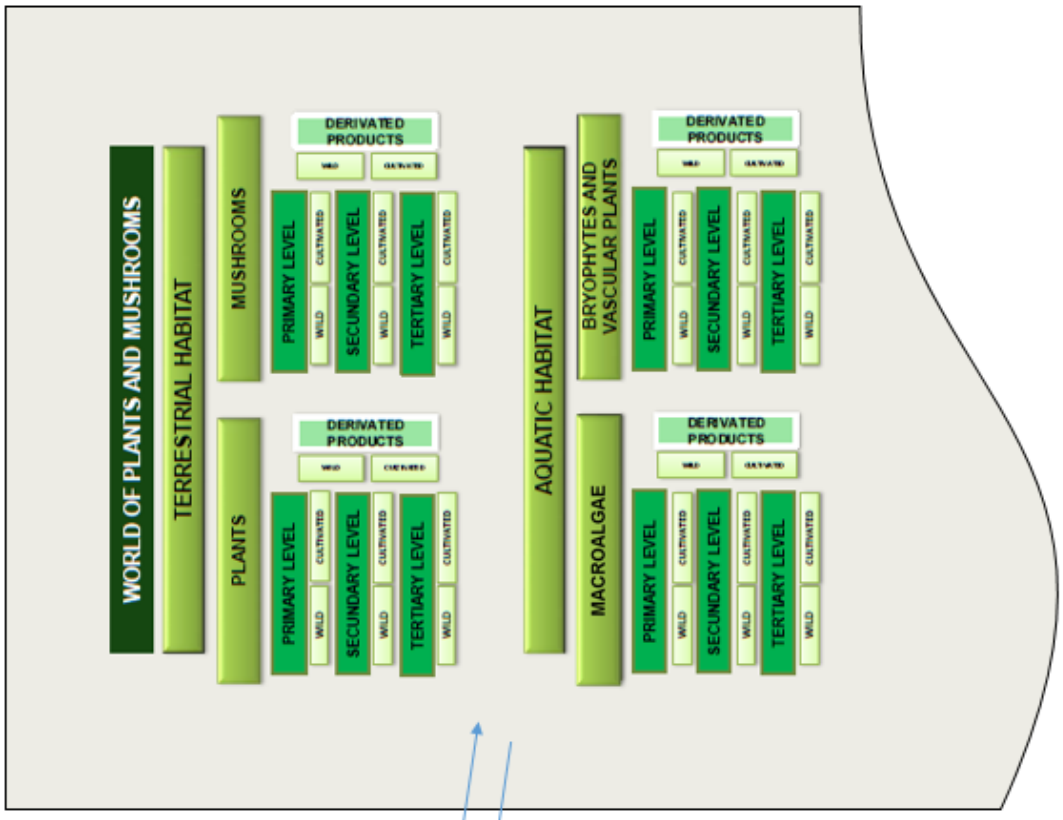
   All the entities in Bullipedia should be linked to one of the eight principal categories, which in turn are divided into smaller subcategories (left side of Figure 17). In the categorization hierarchy proposed by the Bullipedia project we can spot the same semantic relations that we discussed in the section dedicated to WordNet (2.2.1.). General categories and their subcategories are representations of hypernymy-hyponymy relation. For example, in the *Products* category we have five different subcategories. *World of plants and Mushrooms, World of Animals, World of Microrganisms, World of Inorganic Products, World of Elaborated Products* are all hyponyms of *Products.* However, hypernymy-hyponymy is not the only relation that defines Bullipedia category hierarchy. As revealed in the right side of Figure 18, the *World of plants and mushrooms* links to further subcategories. The subcategories of *World of plants and mushrooms* are organized by the relation of meronymy that has three levels: the plant, the part of the plant and the part of the part. Using the example of asparagus, we can say that asparagus plant is the holonym of edible part of the asparagus which in turn is holonym of asparagus trunk.

   In our proposal the link to the ontology is situated at the upper right corner of the page (Figure 15 and 16) and leads us to the general ontology of Bullipedia. From general ontology we can access the ontology page of *World of Plants and Mushrooms.* Given that the ontology of Bullipedia is already previously developed in Bullipedia project we believe that the incompleteness of Wikipedia categorization system can be easily avoided by situating each entity at some level of the ontology.

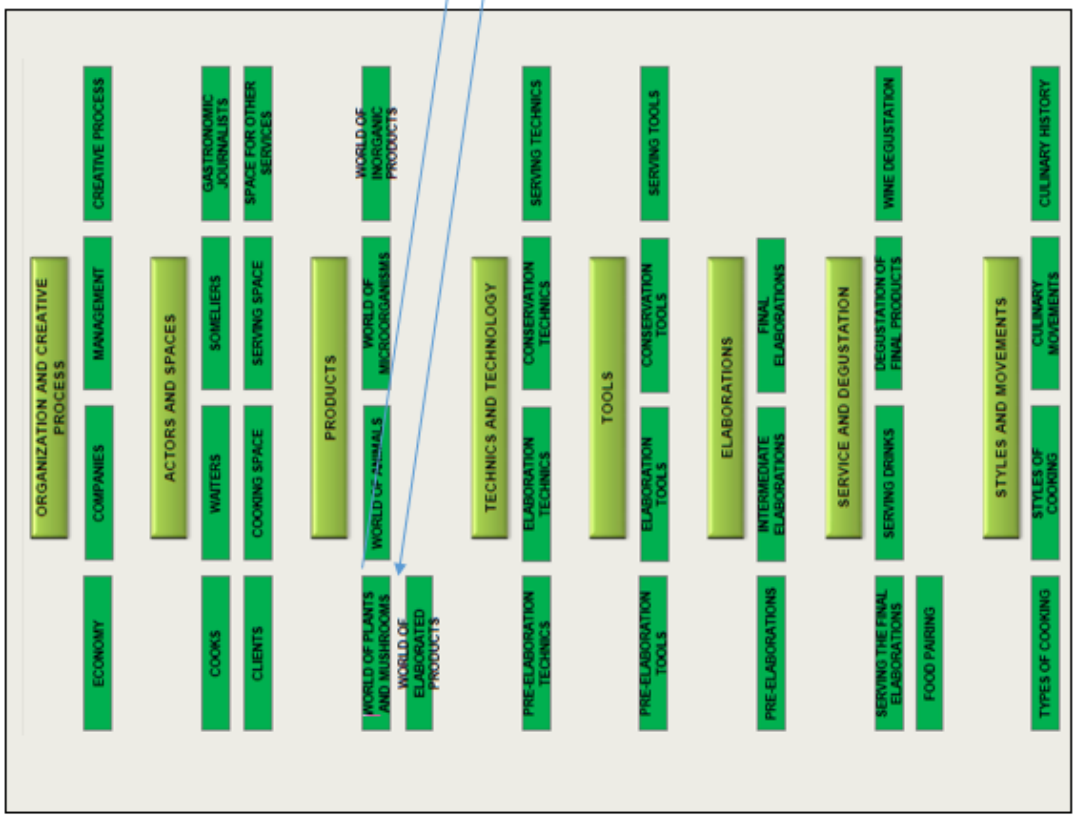
[11] Team of the UB-Bullipedia Unit collaborating in the construction of the classification of gastronomic knowledge: http://www.ub.edu/campusalimentacio/ub-bullipedia/ca/projectes/documentacio.html. Figure 17 shows their proposal.

[12] Team of the UB-Bullipedia Unit collaborating in the construction of the classification of non-elaborated culinary products: http://www.ub.edu/campusalimentacio/ub-bullipedia/ca/projectes/productes.html

Ontology page of the World of Plants and Mushrooms



Ontology page



5. **The information body of Bullipedia should be presented in small extendable boxes.** This way we guarantee a clear visual solution that helps the user to distinguish the knowledge he is looking for and access it directly by clicking on the extendable box (Figure 16). We also take into account that according to Bullipedia project the content of the article in the category of *products* will be previously divided into scientific and gastronomic. We can see in figure 16 how the information content is first divided into two sections: the scientific information and the gastronomic information of asparagus.
6. **Structured data should be preferred to a free text.** That means that the extendable boxes share the same ideals with Wikipedia infoboxes. Using structured data and format of attribute-value ("structured information" boxes in Figure 17) makes it possible to retrieve the information later when it is necessary to infer. It is also recommendable that we established a prototypical pattern for each type of entity. That would help to avoid the problems of incompleteness and inconsistency that each knowledge base should avoid as stated in 2.1.1. and that have been detected in Wikipedia infoboxes (2.3.2. (IV)) and articles. However, we take into account that there is knowledge that has to be presented in format of free text ("free text" boxes in Figure 17) and that the most appropriate format to the properties of the information is always preferred.
7. **Provide links to other resources.** As one of the main characteristics of Bullipedia is the ambition to combine academic, professional and popular knowledge, we believe that external links to WordNet and Wikipedia will complete the scope of Bullipedia. Wikipedia is an unique example of a collaborative tool that aims at containing all the knowledge of the world and more importantly, the Wikipedia knowledge is considered the general belief of the community. Therefore it would complete the point of view of the society in general. WordNet, on the other hand, is a result of years of research and would contribute to the scientific side of Bullipedia.

4. Conclusion

Bullipedia is a unique body of knowledge with particular properties. In order to come up with an adequate knowledge representation model that meets all Bullipedia's needs, we first presented some basic concepts related to our line of research. As there is no resource that would be exactly compatible with Bullpedia we draw inspiration from three different lines of research. First, knowledge representation and reasoning and semantic networks serve as practical examples of how to construct a knowledge base. It is stated that incompleteness and inconsistency of a knowledge base should be avoided at all levels. Second, different possible relations between the entities are defined by the theory of lexical semantics. It should be noted that the semantic relations of hyperonymy/hyponymy and holonymy/meronymy analysed in relation with WordNet are later also detected in Bullipedia ontology. Third, Wikipedia serves as a practical example and point of reference in development of an online encyclopaedia Bullipedia. Based on the overview and possible drawbacks seen in the examples, in chapter 3 we discuss the properties of Bullipedia and develop our model. This work aims to be useful source of information for the development of Bullipedia encyclopaedia. However, this research can also be considered as a source of information to improve the model of Wikipedia. The proposals of this study are of advisory nature and should be viewed as recommendations for the Bullipedia project.

According to our model for Bullipedia, the information should be as structured as possible. Therefore the body of the entry is divided into two. The title and definition (in format of free text) are followed by small extendable boxes with structured information. The boxes of structured information are followed by the boxes with information represented by the free text. The distinction between attributes-values and free text, helps the user to orientate in the "information overload" and distinguish the easily accessible information from the more time-consuming information. General conclusions can be made based on our model:

- All the entities belonging to the same category should be represented by the same template.
- Every entity placed at any level of the hierarchy should be easily accessible in its categorization tree.
- Use of structured information whenever possible makes it possible to retrieve the information later.
- Avoid incompleteness and inconsistency at any level from category hierarchy to the extendable boxes.
- Have external links to other resources like Wikipedia and WordNet and broaden that way the scope of Bullipedia.

Regarding future work, the unique properties of each type of entry (Figure 17) should be analysed. That could be done also by developing a special template for each type of entity. That would be the best way to guarantee complete and consistent content. Another possible theme of research is to analyse the role of meronymy and hypernymy-hyponymy relations as the main organizers of Bullipedia entities. Better mapping of semantic relations in Bullipedia category hierarchy could give interesting results. For example, to analyse the occurrence of different type of meronymy/holonymy relations. Given that Bullipedia will contain structured information, it is possible to extract information from it. It was said above that Wikipedia can be viewed as anything from a corpus, taxonomy, thesaurus, hierarchy of knowledge topics to an ontology and we would like to conclude that the same will also apply to Bullipedia. In other words, we are only in the beginning of the big research that can be done in relation with Bullipedia.

Bibliography:

Adrià, F. (2014). Notes on creativity. Drawing Papers 110. The Drawing Center: New York. Disponible en web [consulted 20.05.2014] http://issuu.com/drawingcenter/docs/drawingpapers110_adria/182

Beckwit, R., Fellbaum, C., Gross, D. and Miller, G.A. (1991). WordNet: A lexical Database Organized on Psycholinguistic Principles.

Brachman, R.J. and Levesque, H.J. (2003). Knowledge Representation and Reasoning, Morgan Kaufmann. Disponible en web [consulted 17.03.2014] http://rair.cogsci.rpi.edu/pai/library/brachmanbook7-17-03.pdf

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C. (2012). Semantic Web Journal.

Malrieu, J. P. (1999). Evaluative Semantics: Cognition, Language, and Ideology, Routledge: London.

Medelyan, O., Milne, D., Legg, C., Witten, I.H. (2009). Mining meaning in Wikipedia. Int. J. Human-Computer Studies 67, 716–754.

Miller, G.A. and Fellbaum, C. (1991). Semantic networks of English. Cognition. 41(1-3), 197-229.

Shapiro, S. C. (1991) Encyclopedia of artificial intelligence Volume 1. A-L,

Vila, M., Rodríguez, H. and Martí, M.A. (2013). Relational paraphrase acquisition from Wikipedia: The WRPA method and corpus. Natural Language Engineering, September 2013, pp 1 – 35.

Acknowledgements

Pere Castells and Unitat UB-Bullipedia in general for structural support and for providing materials and data.

Patricia Benejam for her help in building the example of asparagus.

BullipediaLab and the team of Ferran Adrià in general for letting to participate in the project.