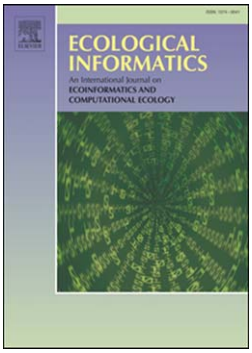
Accepted Manuscript



An authoring tool for decision support systems in context questions of ecological knowledge

Antonio Ferrández, Jesús Peral, Elisa De Gregorio, Juan Trujillo, Alejandro Maté, Luis José Ferrández, Yenory Rojas

PII: S1574-9541(15)00154-5
DOI: doi: 10.1016/j.ecoinf.2015.09.007
Reference: ECOINF 611

To appear in: *Ecological Informatics*

Received date: 9 January 2015
Revised date: 24 June 2015
Accepted date: 1 September 2015

Please cite this article as: Ferrández, Antonio, Peral, Jesús, De Gregorio, Elisa, Trujillo, Juan, Maté, Alejandro, Ferrández, Luis José, Rojas, Yenory, An authoring tool for decision support systems in context questions of ecological knowledge, *Ecological Informatics* (2015), doi: 10.1016/j.ecoinf.2015.09.007

An authoring tool for decision support systems in context questions of ecological knowledge

Antonio Ferrández, Jesús Peral, Elisa De Gregorio, Juan Trujillo, Alejandro Maté, Luis José Ferrández, Yenory Rojas

Dept. Language and Information Systems. Lucentia Research Group. University of Alicante
Carretera San Vicente S/N - Alicante - 03080 - Spain
Phone number:+34-96-590-3400
antonio@dlsi.ua.es, jperal@dlsi.ua.es, edg12@alu.ua.es, jtrujillo@dlsi.ua.es, amate@dlsi.ua.es, ljfp1@alu.ua.es, yrojash@gmail.com

ABSTRACT

Decision support systems (DSS) support business or organizational decision-making activities, which require the access to information that is internally stored in databases or data warehouses, and externally in the Web accessed by Information Retrieval (IR) or Question Answering (QA) systems. Graphical interfaces to query these sources of information ease to constrain dynamically query formulation based on user selections, but they present a lack of flexibility in query formulation, since the expressivity power is reduced to the user interface design. Natural language interfaces (NLI) are expected as the optimal solution. However, especially for non-expert users, a real natural communication is the most difficult to realize effectively.

In this paper, we propose a NLI that improves the interaction between the user and the DSS by means of referencing previous questions or their answers (i.e. anaphora such as the pronoun reference in "What traits are affected by them?"), or by eliding parts of the question (i.e. ellipsis such as "And to glume colour?" after the question "Tell me the QTLs related to awn colour in wheat"). Moreover, in order to overcome one of the main problems of NLIs about the difficulty to adapt a NLI to a new domain, our proposal is based on ontologies that are obtained semi-automatically from a framework that allows the integration of internal and external, structured and unstructured information. Therefore, our proposal can interface with databases, data warehouses, QA and IR systems. Because of the high NL ambiguity of the resolution process, our proposal is presented as an authoring tool that helps the user to query efficiently in natural language. Finally, our proposal is tested on a DSS case scenario about Biotechnology and Agriculture, whose knowledge base is the CEREALAB database as internal structured data, and the Web (e.g. PubMed) as external unstructured information.

Keywords: Decision support system, natural language interface, context questions, ellipsis, anaphora, ontologies

1. Introduction and motivation

Decision support systems (DSS) usually require the access to huge resources of information. As the amount of information available globally on the Web and locally in intranets or databases keeps steadily growing, the necessity of mechanisms for effectively querying this information gains importance at the same pace (Cimiano et al., 2009). Moreover, with the wide availability of smart phones and tablets, the importance of intuitive ways of interacting with electronic devices has grown even more. Natural language interfaces (NLI) are an interesting option to interact with mobile devices due to their limited input and output functionality, which makes graphical interfaces less appealing (Popescu et al., 2003). Clearly, automatic speech recognition is a crucial component towards leveraging the use of NLIs.

Nowadays, the large amount of information obtained by scientific research in Life Sciences is stored in specialized databases (DBs), particularly about Genetics and Biotechnology (Matos et al., 2010). These huge DBs require optimized search strategies in order to extract biological information in a comfortable and efficient way by the user (Jensen, Saric, & Bork, 2006; Altman et al., 2008). In this regard, there is a need to design simple interfaces which work with complex Molecular Biology concepts and ease the biologists the comprehension of the data. As Li et al. (2007) concludes, database query languages (e.g. SQL) can be intimidating to the non-expert, leading to the immense recent popularity for keyword based search or graphical interfaces in spite of their significant limitations. Cimiano et al. (2009) describes different paradigms proposed in the past for querying information collections, among them form filling, query-by-example or menu-based approaches, as well as NLIs, either relying on controlled language or on more or less free language input. Obviously, Natural Language (NL) searching probably constitute the most flexible and effective approach for interrogating a biological DB (Jamil, 2012). Since many biological DBs (e.g. GenBank or UCSC Genome Browser) employ graphical interfaces that are not prepared for arbitrary questioning, previous effort has been made to build NL strategies oriented to Biomedicine and Biotechnology DBs (e.g. Jamil, 2012; Goldsmith et al., 2009; Clegg & Shepherd, 2007; Distelhorst et al., 2003). As Cimiano et al. (2009) states, while the querying paradigm based on NL is generally deemed to be the most intuitive from a usage point of view, it has also been shown to be the most difficult to realize effectively. The main reasons for this difficulty are that:

- NL understanding is indeed a very difficult task due to ambiguities arising at all levels of analysis: morphological, lexical, syntactic, semantic, and pragmatic.
- A reasonably large grammar is required for the system to have an acceptable coverage.
- The NLI needs to be accurate.
- The system should be adaptable to various domains without a significant effort.

Therefore, there is still much work to be done in the field of NLIs. Our proposal and the case study in which it is proved lie on the query of biological knowledge by means of a NLI that handles the context of previous questions and resolves the ellipsis and anaphora ambiguity. For example, in the questions below, the second one needs to resolve the ellipsis to determine the aim of the question: "What QTLs are related to frost tolerance in durum wheat?". The third one needs to resolve the anaphor "these" from the context of the previous questions and their answers.

1. What QTLs are related to frost tolerance in barley?
2. In durum wheat?
3. What other traits are related to these?

The field of Genetic Engineering is an emergent discipline, which has expanded to biomedicine, agriculture and other related domains (Aleksejeva, 2014). Our case study focuses on a Plant Biotechnology industry, whose main target is to create Genetically Modified Organisms (GMOs).

According to the World Health Organization, GMOs are "organisms in which the genetic material (DNA) has been altered in such a way that does not occur naturally" (World Health Organization, 2002) and they are obtained by inserting sequences of DNA from one organism to another. This plant breeding strategy has an important role in the world market, reaching the point that, in 2014, five seed companies control 35% of the global market (Le Buanec, 2008) and 33% of their product are GMOs (Rótolo et al., 2014; ISAAA, 2012; Meijerink & Danse, 2009; ETC, 2008). In cereals, fruits, vegetables, grains and legumes the production of modified seeds keeps growing.

In Plant Biotechnology, the QTL (Quantitative Trait Locus) analysis is highly useful, since allows us to identify the action, interaction, number, and precise location of the chromosomal regions containing one or more genes involved in specific phenotypic features (Falconer & Mackay, 1996; Kearsey, 1998; Lynch & Walsh, 1998; Miles & Wayne, 2008). Thereby, a QTL could be transferred in the laboratory from one organism to another for modifying one or more particular traits.

For this reason, to be updated with QTLs is especially important to design Genetic Engineering protocols, which improve plant varieties (i.e. enhancing its flavor and nutritional value, improving its cold resistance or producing fruits out of season). Given that our knowledge of the function of gene products is increasing rapidly, QTLs databases try to collect all the relations between chromosome positions and biological features of many organisms.

Our proposal facilitates the decision making process because it extends our previous work in Peral et al. (2015), in which, internal structured information (e.g. the CEREALAB database, Milc et al., 2011) and external unstructured data obtained from the Web (e.g. the PubMed URL) are integrated and presented to the user in a dashboard. Here, we extend the NLI in this previous work by handling the context of previous questions and their answers by resolving ellipsis and anaphora. For instance, the user could find in series of questions if a QTL is related to several phenotypical traits (pleiotropy), or if a trait is influenced by several QTLs (polygenic traits with multifactorial inheritance). Moreover, interesting commercial information could be retrieved, like the existence of transgenic varieties and their market price, in order to establish competitive prices for new transgenic seeds.

The paper is structured as follows. In Section 2, we summarize the most relevant related work. In Section 3, we introduce our proposal for the integrated "anaphora + ellipsis + context question" authoring tool for facilitating the decision making process. In Section 4, in order to clarify our proposal, we illustrate the application of our proposal on the case study in which the CEREALAB database is queried. We conclude the paper with the summary of our main contributions and our directions for future works.

2. Related work

As we introduced in our previous motivation section, decision support systems integrate a variety of interfaces for querying databases or data warehouses. So far, these interfaces have been implemented as graphical systems, which ease to dynamically constrain query formulation based on user selections, in order to only build valid questions. However, these graphical interfaces presents the disadvantages of a lack of flexibility in query formulation, since the expressivity power is reduced to the user interface design. Therefore, they provide less expressivity power than textual NL interfaces, as well as they force the user to learn an additional formal language or graphical system. Then, we could consider that NL interfaces are the optimal solution, but they also present some disadvantages such as the itself difficulty of dealing with NL, that is to say the linguistic coverage, ambiguity and the managing of discourse that allows a real natural way of communication (e.g. context questions, anaphora and ellipsis resolution).

The following four subsections discuss the state-of-the-art of these issues: NL interfaces, context questions, anaphora and ellipsis resolution.

2.1.Related work about NL interfaces

In this subsection, we review only those NL interfaces related to our proposal. That is why we group these interfaces according they deal with the problem of anaphora, ellipsis, and context questions, as it is summarized in Table 1. We should emphasize that most of these interfaces do not handle with any of these problems, such as: Popescu et al. (2003), Stratica et al. (2005), or Barbosa et al. (2006).

With regard to those interfaces that faces anaphora problem, the work by Li et al. (2005) presents NaLIX, a generic interactive NL query interface to an XML database, which deals with query pronouns, showing a warning indicating the possible loss of search quality if incorrect anaphora resolution would be made. However, other kinds of anaphors or context questions are not resolved. In Laukaitis & Vasilecas (2007), the authors present an agent based NL dialog architecture for data querying from database management systems. In their architecture, the NL processing module implements morphology, syntax and lexical semantics analysis. The final result of those steps is identified as triplets: entities, relationships and associated probabilities. For this purpose, GATE system (Cunningham et al., 2000) has been used. However, ellipsis or context questions are not handled.

Concerning those interfaces that use ontologies as our proposal does, the work by Cimiano et al. (2008) presents the NLI named ORAKEL. It is an ontology-based NL system: (a) the ontology for a certain knowledge base is used to guide the lexicon construction process; (b) ORAKEL is a NLI which relies on deduction to answer a user's query. The main disadvantage of the system is that the domain lexicon needs to be handcrafted by a domain expert instead of an automatic process that constructs it. The system works on complex sentences, however the following aspects are not resolved: ungrammatical input, unknown words, anaphors, ellipsis, and context questions. The proposal by Jamil (2012) is remarkable since he presents a generic NL plug-in for querying biological databases, close to the topic of our case study. He proposes a method to map NL questions to semantically equivalent database specific SQL queries. Although the plug-in introduced is generic and facilitates connecting user selected NLI to arbitrary databases, it does not resolve ellipsis or anaphora, nor context questions.

Regarding the interfaces that deal with context questions, the proposal by Elhai et al. (2009), BioBIKE, is presented as a graphical programming interface that allows the interaction of biologists with information of interest to them (the biological knowledge base). The main problem with this approach is the rigidity of the questions that have to be posed in the specific language of the environment. However, it is possible to construct a progressive series of queries (using BioBIKE language), each one utilizing on the result of the previous. This is similar to simple context questions but the user is responsible for translating from NL to the specific language of BioBIKE. It does not resolve ellipsis or anaphora. In Distelhorst et al. (2003), the paper describes a constrained NLI Plug-in to a large knowledge base, the Foundational Model of Anatomy (FMA). The interface, called GAPP, handles simple or nested questions that can be parsed to the form, subject-relation-object. The system processes nested questions, such as "Which part of the thorax contains the lung?". This question is modelled in GAPP by performing the internal query first ("Which part of the thorax?"). The output of the internal query then serves as the input for the top level query. However they do not deal with ellipsis or anaphora.

Finally, an interesting proposal is the one by Goldsmith et al. (2008), which faces the access to databases from the Information Retrieval (IR) point of view (our proposal can interact with IR and Question Answering systems). They present CSIR (Cognition Search Information Retrieval), a Natural

Language Processing (NLP) technology that improves access to the MEDLINE database of scientific abstracts. The main problem of this approach is that the questions have a very simple structure (most of them phrases) and the system does not address questions with complex structure and NLP problems.

System	Lexical Analysis	Syntactic Analysis	Semantic Analysis	Ellipsis Res.	Anaphora Res.	Ontology/ Semantics	Specific Domain	Context Questions
Popescu et al.	Yes	No	No	No	No	Wordnet	No	No
Stratica et al.	Yes	Yes	Yes	No	No	Wordnet	Yes	No
Barbosa et al.	Yes	Yes	No	No	No	No	No	No
NaLIX	Yes	Yes	Yes	No	Yes	No	No	No
Laukaitis & Vasilecas	Yes	Yes	Yes	No	Yes	No	No	No
ORAKEL	Yes	Yes	No	No	No	Yes	Yes	No
Jamil	Yes	Yes	Yes	No	No	Yes	Yes	No
BioBIKE	No	No	No	No	No	No	Yes	Yes
Distelhorst et al.	Yes	Yes	Yes	No	No	Own / Wordnet	Yes	Yes

Table 1. Comparative table between NLIs

As Table 1 shows, we can conclude that most of these NLIs do not deal with an integrated anaphora and ellipsis resolution, and only some of them manage context questions but with limited solutions (e.g. rigidity of the questions posed in the specific language of the system). Moreover, our proposal is based on ontologies semi-automatically generated, which ease the portability process of the NLI to various domains without a significant effort.

2.2.Related work about context questions

Rarely questions are asked in isolation because it is usual that a user may have follow-up questions requiring additional information or clarifying the searched information, which are usually called *questions in context* or *context questions*. Remarkable initiatives to address these context questions are the TREC (Voorhees, 2001) and NTCIR (Fukumoto et al., 2003, 2004, 2007) Question Answering (QA) track competitions , and the Question Answering on Speech Transcriptions (QAST) CLEF track (Lamel et al., 2008). One of their tasks had to deal with questions posed in the context of previous questions and answers through series of questions.

An outstanding work in these tasks was the one by Harabagiu et al. (2001) that encoded an efficient way of modelling context via reference resolution. This work runs a coreference resolution process prior to the recognition of the expected answer type. They stated that the resolution of the following forms of reference is required:

- Demonstrative pronouns (“On what day did this happen?”);
- Third person pronouns (“What California winery does he own?”);
- Possessive pronouns (“What was his first radio song?”);
- Definite nominal/descriptions (“What executive from the company was a member of the Supreme Council in 1994?”);
- Nominalizations of verbs (“In what facility was it constructed? When was construction begun?”);

- Elliptical reference ("How many species of spiders are there? How many Ø are poisonous to humans?");
- Meronymic reference ("Which museum in Florence was damaged by a major bomb explosion in 1993? Which galleries were involved?").

Their reference resolution algorithm is different from reference resolution algorithms used in discourse or dialog processing, because their aim is to identify the question that either contains the antecedent (i.e. the entity to be referred) of the reference (also called anaphor, such as a pronoun) or expects an answer that contains the antecedent. Therefore, they do not aim to resolve the reference. Once they conclude that a previous question or its answer can contain the antecedent of a reference, they only combine the keywords of both questions. Instead of resolving references using discourse information, this system first identifies the questions that contain the potential referents and uses those questions and the current question to identify the target paragraph.

The TREC 2004 Question Answering track (Voorhees, 2005) varied the context questions in the way that each series of questions were related to a single target. Most of the system participants in this competition simply appended the target to the question. Another common approach was to replace all pronouns in the questions with the target. While many (but not all) pronouns in the questions did in fact refer to the target, this approach suffered when the question used a definite noun phrase rather than a pronoun to refer to the target (e.g., using "the band" when the target of the series was Nirvana). Finally, other systems tried varying degrees of true anaphora resolution to resolve appropriately references in the questions. It is difficult to judge how much benefit these systems received from this more extensive processing since the majority of pronoun references were to the target.

Similarly, in the CLEF 2007 Multilingual Question Answering Track (Giampiccolo et al., 2007), topic-related questions track consisted of clusters of questions, which were related to the same topic. This was accomplished either by co-reference either by anaphoric reference to the topic declared implicitly in the first question or in its answer. They observed an overall decrease in the accuracy reached by the systems when treating linked questions, which is mainly due to the non-handling of anaphora.

In Kato et al. (2004, 2005), NTCIR4/NTCIR5 - QAC Subtask 3 is described as a challenge to measure objectively and quantitatively the ability of QA systems to interactively participate in dialogues for accessing information, such as when gathering information for a report on a specific topic, or when browsing information of interest to the user. Therefore, this series of questions and the answers to those questions comprise an information access dialogue. Although systems are supposed to participate in dialogue interactively, the interaction is only simulated; systems answer a series of questions in a batch mode. They conclude that the participants' techniques employed for context processing are rather simple. These techniques provide a basis for further development. In the most prevailing case, systems do not analyze referential expressions in a given question at all, but simply treat that question as a continuation of preceding questions. Systems that use keywords extracted at the question analysis stage in the subsequent stages take keywords in the preceding questions into consideration in addition to those in the current one. Other approaches that intensively use NLP have not achieved satisfactory results. For instance, Matsuda & Fukumoto (2005) categorizes reference expressions into three types: pronouns, zero anaphora of a case element of verbs, and zero anaphora of a modifier of nouns. The latter two types are processed using case frames and co-occurrence data from the EDR Japanese Co-occurrence Dictionary, respectively. This reference resolution is based on application-independent linguistic knowledge and linguistic analysis, and attempts to address a wide range of phenomena.

As conclusion, we can observe that most of the approaches face context questions by means of naïve anaphora resolution, and ellipsis resolution is afforded by just adding previous keyword questions.

2.3.Related work about anaphora resolution

In this subsection, we detail the anaphora resolution work related to context questions. We can distinguish between traditional approaches, and those that are specialized in context questions, mainly through discourse modelling. The latter are more important for the issue in this paper, in which the discourse research mainly addresses two important issues (Sun & Chai, 2007): (1) what information is to be captured from the discourse; and (2) how such information can be represented for language interpretation and generation. Many theories have been developed for both issues, such as Hobbs theory (Hobbs, 1985), Rhetorical Structure Theory (Mann & Thompson, 1987) and Centering Theory (Grosz et al., 1995); and theories for dialogues, such as Grosz & Sidner (1986) conversation theory and Discourse Representation Theory (Mann & Thompson, 1987).

In Chai & Jin (2004), the authors discuss the role of discourse modeling in context QA. They state that every question and its answer have a discourse status with respect to an entire QA session. This discourse status includes two aspects. The first aspect relates to discourse roles of entities in a question and the corresponding answer. Entities (such as noun phrase, verb phrase, preposition phrase, etc.) in a question carry distinctive roles that indicate what is the topic or focus of a question in terms of the overall information seeking discourse. Topic relates to the "aboutness" of a question and focus relates to a specific perspective of the topic. The second aspect of discourse status relates to discourse transitions that indicate how discourse roles are changed from one question to another as the interaction proceeds and how such changes reflect the progress of user information needs.

Sun & Chai (2007) investigates the role of discourse processing and its implication on question expansion for a sequence of questions on the document retrieval task previous to QA. They examine three models driven by Centering Theory (Grosz et al., 1995) for discourse processing: (1) a reference model that resolves pronoun references for each question; (2) a forward model that adds question terms from the previous question based on its forward looking centers; and (3) a transition model that selectively adds question terms according to the transitions identified between adjacent questions. The adding of question terms is performed through a set of heuristics based on linguistic knowledge, which occurs when the Centering Theory detects a change in the focus.

Regarding traditional approaches for anaphora resolution (Zheng et al., 2011), as well as typical pronominal resolution (Palomar et al., 2001, Bellot et al., 2002), definite descriptions should be resolved as previous subsection stated. In Vieira & Poesio (2000), the authors conclude that definite descriptions are not primarily anaphoric (descriptions that denote the same discourse entity as their antecedent), because half of the time they are used to introduce a new entity in the discourse. Their heuristic to determine the anaphoric references is when synonymy (a house vs the home), generalization/hyperonymy (an oak vs the tree), specialization/hyponymy (a tree vs the oak) or meronymy semantic relations are detected. Another heuristic is when both description and antecedent have the same head noun and compatible modifiers (a blue house vs the house), although it is not satisfied always, for example when a proper name is used (Bill Clinton vs the president). The authors stablish several additional heuristics. For example, the one that states that definite descriptions in appositive (Glenn Cox, the president of Phillips Petroleum Co.) and copular constructions (the man most likely to gain custody of all this is a career politician named David Dinkins) tend to be discourse-new, related to the NP to which they are attached.

2.4.Related work about ellipsis

In Carbonell (1983), discourse phenomena that occur frequently in task oriented man machine dialogs is reviewed, demonstrating the necessity of handling ellipsis, anaphora, and other abbreviatory

devices in order to achieve convivial user interaction. The author concludes that users prefer to generate terse or fragmentary utterances instead of longer, more complete "stand-alone" expressions, even when given clear instructions to the contrary. This author also distinguishes between syntactic and semantic ellipsis, where the latter refer to ellipsed information not manifested as syntactically incomplete structures, but as semantically incomplete propositions. The method proposed to resolve both kinds of ellipsis is based on a semantic case-frame approach and ad-hoc contextual substitution rules.

In Díaz de Ilarraza et al. (1990), the authors detail the two major subproblems of ellipsis resolution in dialogued systems: (1) the analysis of the elliptical sentence; (2) the reconstruction of the elided fragments. They face the resolution by means of detecting the missing of mandatory elements in the question, both in syntactical and conceptual ellipsis. When values of mandatory descriptors are not present, their system will generate expectations for instances that could fill the descriptors. Regarding syntactical ellipsis, they distinguish between (1) substitution of a syntactic category by another, or (2) Adjunction of modifiers chains to a central one.

The work by Williams, S. (2000) describes the anaphoric reference and ellipsis resolution component of a Spoken Language System, which provides telephone-based access to email. This faces the resolution by storing a sorted list of utterances that appear in previous questions. In this list, a set of domain specific entities are also stored, such as "message, email, header or folder". These specific entities are searched first, and if they are present in the question to resolve, then the remaining utterances are searched.

The work by Tomioka, S. (2008) focuses on elided verbal phrases (VP), which demand the presence of the "same" VP or VP meaning. But, as the author concludes, defining the "sameness" is no easy task. Therefore, he proposes a meaning recovery strategy similar to anaphora resolution.

As conclusion, an extensive coverage of anaphora and ellipsis resolution approaches has been developed. The contributions of our proposal in comparison to the state-of-the-art can be summed up in the following points: i) our proposal is modular in order to allow the use of different linguistic tools; ii) in the implementation of our proposal proved in a case study, we deal with anaphora (pronominal and definite descriptions) and ellipsis (syntactic ellipsis supported by semantic compatibility) in an interrelated way; ii) it offers an integrated architecture that interfaces to different structured and unstructured sources.

3. A novel authoring tool that handles the context of questions and facilitates the DSS

In this section, our proposal is fully explained in two subsections. The first subsection describes the architecture in which our proposal is integrated, which accesses structured and unstructured information in order to facilitate the decision making process. The second subsection details the integrated solution to deal with anaphora, ellipsis and context questions, and it also presents an example of the application of our proposal, which is based on the case study that will be developed in the following section. This case study deals with the questions posed by a plant breeder enterprise in order to carry out new breeding programs experimenting with the new advances in Genetics. The questions will query both internal structured and external unstructured information.

3.1.The integration of our proposal into an architecture to access structured and unstructured information

Our proposal is embedded into our previous work in Peral et al. (2015) that provides a framework for integrating unstructured and structured information in a common interface, like a dashboard, where easily the user can interpret his needed information. This framework is depicted in Figure 1. In the GUI

(Graphical User Interface) module of this Figure, the user or decision maker poses a NL question and selects the sources to be searched such as a specific database or data warehouse, or in a specific QA domain. In this way, our proposal can deal with a non-limited number of sources, which are represented by the "Node n" in this Figure. The GUI module passes the NL question to the Distributor/Integrator module that also sends it to the set of selected nodes, such as the Data Warehouse (DW) and Question Answering (QA) nodes. Each specialized node disposes of the proper interface in order to process adequately the NL question and to produce the suitable output information. Then, the Distributor/Integrator coordinates the running of each specialized node, gathering the output of these nodes in order to send the fused information to the GUI module. Finally, the GUI is responsible for displaying the results as a dashboard, which integrates both external and internal data.

During the setup phase of this framework, the source nodes that contain the required information to be searched are prepared by creating the corresponding ontologies. For example, the first time that a specific DW is connected to the framework, the DW ontology that describes the DW scheme is created and mapped with the remaining node ontologies, which will allow its integration with the remaining nodes connected to the framework. The process to generate semi-automatically the DW ontology is based on the proposal by Santoso et al. (2010). We use the knowledge obtained in the creation of the conceptual model of the DW. This is performed in five steps. In the first step, a concept is created automatically in the ontology for each dimension level existing in the conceptual model. In the second step, a concept representing the DW is related to each of the finest grained levels on each dimension. During the third step, each drill down relationship in the conceptual model is transformed into a *has_members* relationship in the ontology. Conversely, each roll up relationship is transformed into an *is_a_member_of* relationship between ontological concepts. Finally, the fourth step is manual in which the ontology is enriched with additional semantics and relationships. The designer may change the semantics of some of the automatically generated relationships, modifying or renaming the relationships created by default, specifying relationships across different dimensions or including attributes that are considered useful for querying purposes. To conclude, in the fifth step the ontology is populated automatically with all the information stored in the DW where each data is stored as an instance of the corresponding node. This ontology will be used as the knowledge base in the following scheme to deal with context questions through resolving anaphora and ellipsis.

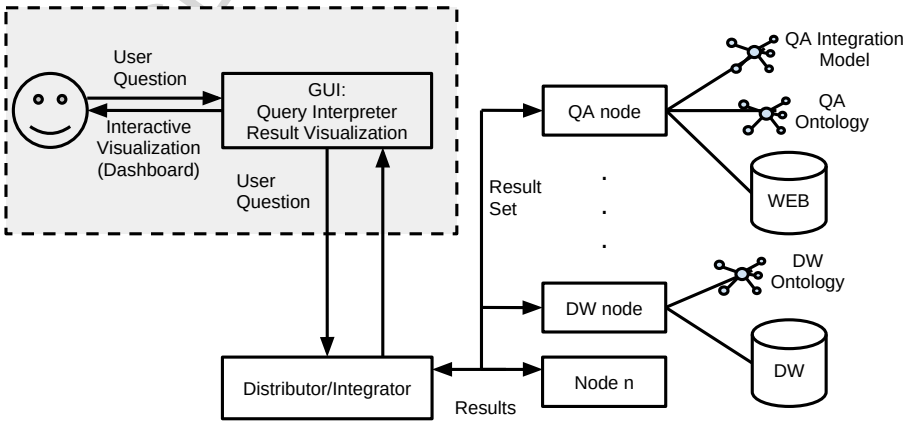


Figure 1. Framework for integrating unstructured and structured information (Peral et al., 2015)

3.2.The integrated solution to deal with anaphora, ellipsis and context questions

In this subsection, we detail our proposal to resolve anaphora, ellipsis and context questions, and we describe how it is embedded in the shaded part of Figure 1. Our proposal is depicted in Figure 2, in which firstly, the GUI element interacts with the user through our NLI (Llopis & Ferrández, 2012) in order to generate the "text user question" in Figure 2. This NLI is used as a query-authoring service that has pre-programmed elements for the development of interactive multimedia help in the query construction, such as syntax colouring, text completions or keyword highlighting. It improves the system usability allowing the decision maker to early detect errors in questions by automatically distinguishing between linguistic (e.g. errors due to lexical or syntactic mistakes) and conceptual failures (e.g. errors due to the lack of a specific relation between tables in the database).

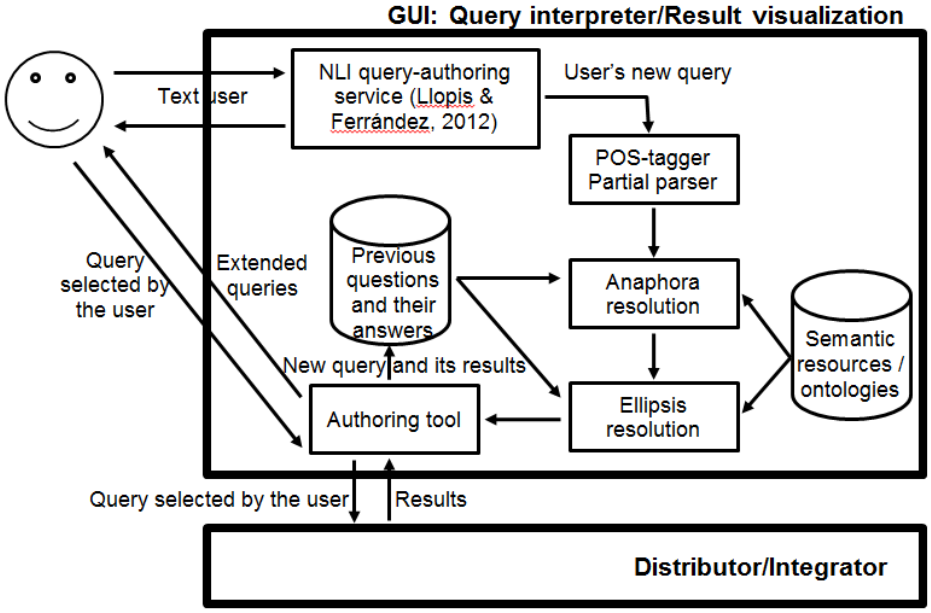


Figure 2. Architecture for resolving anaphora, ellipsis and context questions

We have extended this NLI by dealing with anaphora, ellipsis and context questions. This extension is achieved by receiving the "user's new query" (Figure 2), which is POS-tagged (it returns lemma and lexical information of each word) and partial parsed (it returns syntactic trees). A data structure that is explained next will store all the lexical (obtained from a POS-tagger), syntactical (from a partial parser) and semantical knowledge (from the semantic resources and ontologies) required by the anaphora and ellipsis resolution modules. These resolution modules will handle the context stored in the "previous questions and their answers" database (Figure 2). They will generate a set of "extended queries" (Figure 2), with which the authoring tool module will interact with the user in order to help him/her to select the "query selected by the user" (examples of this interaction are presented in Figure 3 and Figure 5). This selected query will be passed to the Distributor/Integrator module, which is in charge of sending it to the selected nodes (e.g. the DW and QA nodes), gathering the output of these nodes, and sending the output to GUI module. Finally, the GUI module will display the results as a dashboard (examples in Figure 8 and Figure 9), and it will update the "previous questions and their answers" database (Figure 2).

The proposed algorithm is presented in Algorithm 1, in which the data structures used for the interaction between modules are specified. Firstly, the scheme receives the "user's new query" (see Figure 2) posed by the user (*NLQ* in Algorithm 1), which is a string data type (i.e. free NL text). It also receives: the ontology (*Ont*) semi-automatically generated as explained in subsection 3.1; the semantic resources shared by all the nodes (*WN*, e.g. WordNet); and the list of previous questions posed by the user (*previous_questions*) and their corresponding answers (*previous_answers*).

```
ALGORITHM  anaphora_ellipsis_context_questions_resolution

INPUT      NLQ: string; // It stands for the variable NLQ, which is of type "string"
           Ont: Ontology;
           WN: Semantic_resources;    // For example WordNet
           previous_questions: list_of_lists_of_slot_structures;
           previous_answers: list_of_lists_of_slot_structures;

OUTPUT     NLQ_concepts: list_of_lists_of_slot_structures;

VAR        parsedSentence: list_of_slot_structures;

BEGIN

   //Partial parsing of the question
   parsedSentence = partialParsing(NLQ);
   //Mapping of phrases into the ontology and semantic resources
   NLQ_concepts = semantic_enrichment(parsedSentence, Ont, WN);

   // Anaphora detection and resolution
   NLQ_concepts = anaphora_resolution(NLQ_concepts, previous_questions,
         previous_answers, Ont, WN);

   // Ellipsis detection and resolution
   NLQ_concepts = ellipsis_resolution(NLQ_concepts, previous_questions, Ont, WN);

END ALGORITHM
```

Algorithm 1. Integrated algorithm to handle anaphora, ellipsis and context questions

The NLQ question is partial parsed, where noun phrases (NP), prepositional phrases (PP) and verbal phrases (VP) are fully parsed (i.e. the NPs can have nested structures such as PPs, appositions, relative clauses or coordinated NPs, e.g. "frost tolerance in wheat"), and the chunks not included in these phrases are skipped in the parsing. These phrases represent the main concepts involved in the question. For example, consider the question: "What QTLs are related to frost tolerance in wheat?" The NPs extracted are "QTLs" and "frost tolerance"; the PP "in wheat"; the VP is "are related to"; and the chunks not included in these phrases "what" and "?". The list *parsedSentence* in Algorithm 1 stores these phrases in the sequential order that they appear in the question. This list contains *slot structures* (Ferrández et al., 1999) extracted from the parser, which store the morphological knowledge (in the structure "conc", such as number and gender), an identifier (marked in the following examples as upper cases such as X), syntactic knowledge (e.g. the slot structures of nested phrases), the lemma and the question term of each question word. In the following examples, the slot structures of nested phrases and lemmas are not shown with the aim of clarity. For the above mentioned query "What QTLs are related to frost tolerance in

wheat?", parsedSentence stays as follows (from now on, the list structure will be denoted between square brackets):

```
[
  what
  np (conc (properName, plural), X, n(QTLs))
  vp (conc (plural), Y, v(are), v(related), prep(to))
  np (conc (commonName, sing), Z, n(frost), n(tolerance), T)
  pp (conc (), T, prep(in), U)
  np (conc (commonName, sing), U, n(wheat))
  ?
]
```

Then, these slot structures (parsedSentence) are extended (*semantic_enrichment* in Algorithm 1) with semantic information of the ontology (*Ont* that is generated as explained in subsection 3.1) and external semantic sources (*WN* e.g. WordNet). The extended slot structure is stored in *NLQ_concepts*, which is a *list_of_lists_of_slot_structures* given that it will store each suggestion of reformulation of the question (e.g. one reformulation for each likely solution for an anaphor). The "authoring tool" module (Figure 2) will allow the user to select one of these suggestions. The slot structures that are found in the ontology will be the foremost concepts used for the following search process. For example, "QTL" matches with the level "QTL" in the Ont used in our case study (see Figure 7); "frost tolerance" is matched as an instance of the level "Trait" in Ont; and "wheat" is matched as an instance of the level "Species" of plants. Moreover, when an exact matching with Ont is not reached, each partial matching is suggested as new reformulations of the question. For example, in the question "I would like to know the QTLs related to rust in wheat", the concept "rust" partial matches with three instances of "Trait": "resistance to leaf rust", "resistance to stem rust", and "resistance to stripe rust". Therefore, a question for each "Trait" is suggested to the user in the authoring tool as possible reformulations: "I would like to know the QTLs related to resistance to leaf rust in wheat.", "I would like to know the QTLs related to resistance to stem rust in wheat.", "I would like to know the QTLs related to resistance to stripe rust in wheat.".

Next, each slot structure is extended by means of WN (e.g. using synonymy, hyperonymy, hyponymy and the remaining relations of WordNet). For instance, the term "tolerance" is linked with its WordNet synset with all its synonyms (e.g. permissiveness or allowance). These synonyms will be used to tag the slot structure with additional ontology concepts. For example, in the question "Which genes influence the resistance to plagues in wheat?", the concept "plagues" is not found in the ontology. But after the WordNet tagging, a synonym of "plague" is "group/swarm of insects", which is a "Trait" of the wheat, and the Trait instances "resistance to Russian wheat aphid" and "resistance to hessian fly" refer to plagues because "aphid" and "fly" are insects. Finally, a question for each "Trait" is suggested to the user in the authoring tool: "Which genes influence the resistance to Russian wheat aphid in wheat?", "Which genes influence the resistance to hessian fly in wheat?".

The semantic extension of each concept is stored as two additional arguments of the "conc" structure. For example, the concept "wheat" would have the following "conc" structure: np (conc (commonName, sing, ***[Species], [12142085, 07803545]***), U, n(wheat)). These arguments are lists in order to store several likely matchings sorted by probability of certainty. The first additional argument stands for a matching with the ontology Ont, whereas the second additional argument represents the list with the synsets related to the concept (synonym, hyperonym, etc.) in WN. In case that a matching is not obtained, the symbol "_" is stored, as it occurs in: np (conc (properName, plural, ***[QTL], _***), X, n(QTLs)), which means that QTL is not found in WN.

After the generation of the NLQ_concepts for the present question, the "anaphora resolution" module (*anaphora_resolution*) is run (see Algorithm 2). It receives as input the NLQ_concepts, previous_questions and previous_answers (a similar list of lists of slot structures with all these information referred to previous questions and their answers, which corresponds to "Previous questions and their answers" database in Figure 2). All these lists have been extended with Ont and WN, as previously described.

The *antecedent_extraction* process in Algorithm 2 pre-process the NLQ_concepts, previous_questions and previous_answers lists, in order to index all the slot structures of the possible solutions of an anaphor (e.g. the noun phrases for pronominal anaphors). These are stored in the *list_antec* structure, which also groups them by different questions. The *anaphor(NLQ_concepts)* process performs the search of anaphors in NLQ_concepts (the present question) through the lexical (e.g. pron(conc (thirdPerson), …) for a pronominal anaphor) and syntactical information (e.g. for a definite description: np(conc(…), …, determiner(the), n(tolerance))).

The resolution is carried out by the substitution of the anaphor for the list of slot structures of the possible solutions or antecedents to which it could refer (sorted by probability of certainty). In this way new reformulations of the question are generated in *generation_of_reformulations*. Whatever anaphora resolution scheme can be used in this stage. For instance, in the case study developed in the following section, we are using our pronoun resolution proposal Palomar et al. (2001) and our definite description resolution Muñoz et al. (2000), but in the future we plan to comparatively use other state-of-the-art coreference resolution systems (e.g. Berkeley's Durrett & Klein, 2013). Our pronoun and definite description resolution module is summarized in Algorithm 2, which will use the morphological, syntactical and semantical knowledge stored in the slot structures of the previous questions posed by the user (previous_questions) and their corresponding answers (previous_answers). It sorts the set of possible antecedents or solutions for each anaphor (list_antec) by distinguishing between constraints and preferences that will be applied to the different candidate antecedents. Each type of anaphor have a set of constraints (e.g. for pronouns, *anaphor.constraint()* will be number/gender agreement and c-command constraints) and preferences (e.g. for definite descriptions, *anaphor.preference()* will be heuristics such as noun phrase head matching or noun phrase modifiers matching).

```
ALGORITHM  anaphora_resolution

INPUT      NLQ_concepts: list_of_lists_of_slot_structures;
           previous_questions: list_of_lists_of_slot_structures;
           previous_answers: list_of_lists_of_slot_structures;
           Ont: Ontology;
           WN: Semantic_resources;

OUTPUT     NLQ_concepts: list_of_lists_of_slot_structures;

VAR        list_antec: list_of_list_of_indexes_of_slot_structures;

BEGIN

   list_antec = antecedent_extraction(  NLQ_concepts, previous_questions,
                                        previous_answers );

   For each anaphor(NLQ_concepts) do                    // Anaphora detection
      For each anaphor.constraint () do                 // Constraint sorting
        list_antec = anaphor.constraint (NLQ_concepts, list_antec, Ont, WN);
      For each anaphor.preference() do                  // Preference sorting
        list_antec = anaphor.preference (NLQ_concepts, list_antec, Ont, WN);
   End for

   NLQ_concepts = generation_of_reformulations(list_antec);

END ALGORITHM
```

Algorithm 2. Anaphora resolution algorithm

Subsequently, ellipsis (*ellipsis_resolution* in Algorithm 3) is resolved in a similar way using the NLQ_concepts extended with the reformulations of the anaphora resolution module, previous_questions and previous_answers. The resolution is performed through the parallelism with the previous question and the QA traditional processing of the question. With the QA question processing, we mean that a traditional analysis is performed, in which there are compulsory terms that should appear in the question, such as a VP, and optional but categorized phrases such as the interrogative particle with its following NP that usually restricts the type of the searched information, and additional phrases. With regard to the parallelism, the slot structures are paired with those of the same syntactic and semantic category. In case of missing phrase slot structures, the corresponding pairs are proposed as possible reformulations of the question.

```
ALGORITHM  ellipsis_resolution

INPUT      NLQ_concepts: list_of_lists_of_slot_structures;
           previous_questions: list_of_lists_of_slot_structures;
           previous_answers: list_of_lists_of_slot_structures;
           Ont: Ontology;
           WN: Semantic_resources;

OUTPUT     NLQ_concepts: list_of_lists_of_slot_structures;

VAR        list_antec: list_of_list_of_slot_structures;

BEGIN

   If not QA_complete_sentence(NLQ_concepts)
      For each phrase(NLQ_concepts, previous_questions) do
         If previous_questions.phrase isA particle
            and NLQ_concepts.phrase isNotA particle
                  add_phrase(NLQ_concepts, previous_questions.phrase);
         If previous_questions.phrase isA vp
            and NLQ_concepts.phrase isNotA vp
                  add_phrase(NLQ_concepts, previous_questions.phrase);
         If previous_questions.phrase isA np
            and NLQ_concepts.phrase isNotA np
            and not compatible(NLQ_concepts.phrase, previous_questions.phrase)
                  add_phrase(NLQ_concepts, previous_questions.phrase);
         If previous_questions.phrase isA pp
            and NLQ_concepts.phrase isNotA pp
                  add_phrase(NLQ_concepts, previous_questions.phrase);
      End for
   End if

END ALGORITHM
```

Algorithm 3. Ellipsis resolution algorithm

For instance, let us suppose the previously mentioned question "What QTLs are related to frost tolerance in wheat?". The method *QA_complete_sentence* in Algorithm 3 will return true, so there is not ellipsis resolution, because in the QA question processing (Ferrández et al., 2009), a well formed question of type *entity_group – QTL* is detected (i.e. the syntactic-semantic pattern is matched: "[What] + [NP of QTL type] + [VP] + [set of phrases]?"). But if this question is followed by an user's question such as "In barley?", then QA_complete_sentence would return false because of the missing of the compulsory syntactic structure VP that avoids the QA question processing to classify it. The system detects "barley" as an instance of "Species". Therefore, the following slot structures are stored in NLQ_concepts:

```
[ pp (conc (_, _), V, prep(in), W)
  np (conc (commonName, sing, [Species], [07803093, 12123244]), W, n(barley))
  ? ]
```

Then, the line "For each phrase(NLQ_concepts, previous_questions)" in Algorithm 3 will traverse in parallel the sets of particles, verbal, noun and prepositional phrases. The missing syntactic structure VP is

restored from the previous question, similarly to the interrogative particle "what" and its following NP that delimits the aim of the question search ("QTL"). With regard to the remaining elided slot structures, in this example there is the ambiguity between restoring "np (conc (commonName, sing, [Trait], [05033410]), Z, n(frost), n(tolerance), T)" or "pp (conc (_, _), T, prep(in), U); np (conc (commonName, sing, [Species], [12142085, 07803545]), U, n(wheat))". This ambiguity is solved to the former due to the matching of the syntactical structure (PP with preposition "in") and the semantic category (process *compatible*, which succeeds when an exact matching or semantic matching, such as hypernym or synonym, between the semantic lists occurs, as in this case: [Species]). Therefore, the final NLQ_concepts remains as the following, where the "*" in the identifiers marks the added phrases (this scheme is also used for the solutions of an anaphor), in order to be confirmed by the user through the authoring tool. The whole slot structure is not presented to the user, since just the question words are printed as it is presented in Figure 3:

```
[ what (*)
  np (conc (properName, plural, [QTL], _), X*, n(QTLs))
  vp (conc (plural, _, [00713167, 02724417, 02458103]), Y*, v(are), v(related),
prep(to))
  np (conc (commonName, sing, [Trait], [05033410]), Z*, n(frost), n(tolerance), T)
  pp (conc (_, _), V, prep(in), W)
  np (conc (commonName, sing, [Species], [07803093, 12123244]), W, n(barley))
  ? (*) ]
```

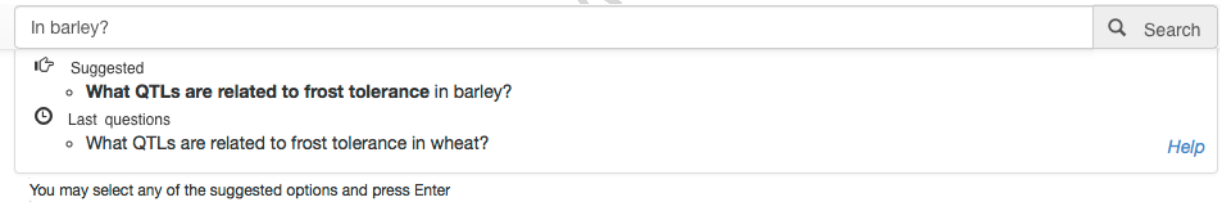


Figure 3. Screenshot of the context questions "In barley?" that follows "What QTLs are related to frost tolerance in wheat?"

With regard to anaphora resolution, let us suppose that after that, the user poses the following question: Which traits are affected by them? That is to say, the user wants to know the traits affected by each QTL answered in the second question. Then, NLQ_concepts is obtained as:

```
[ which
  np (conc (commonName, plural, [Trait], [04616059]), A, n(traits))
  vp (conc (plural, _, [00137313, 00019448, 02677097]), B, v(are), v(affected))
  pp (conc (_, _), C, prep(by), D)
  pron (conc (thirdPerson, plural, _, _), D, them)
  ? ]
```

After these two questions, the previous_questions and previous_answers structures contain the slot structures in Figure 4. They are a list of lists (*list_of_lists_of_slot_structures* in Algorithm 1), where each list item is a list of slot structures of each question. In other words, previous_questions = [[slot structures of question 1], [slot structures of question 2]]. These lists allow the access to all the knowledge about each question.

Next, the anaphoric expression is detected by the slot structure: pron(conc (thirdPerson, plural, …), …)). The list of possible solutions or antecedents is only formed by the only plural phrase "QTLs". Since this solution is the NP that delimits the aim of the question search and follows the interrogative particle in the second question, list of solutions is formed by the slot structure of QTLs, and the slot structures of the answers of the second questions previously stored in *previous_answers* ("QWsv.DiMo-5H.1" and "QWsv.DiMo-5H.2"). A new question for each possible solution is presented to the user through the authoring tool: "Which traits are affected by QTLs?", "Which traits are affected by QWsv.DiMo-5H.1?" and "Which traits are affected by QWsv.DiMo-5H.2?". Each reformulation is stored in NLQ_concepts as a list of sentences:

```
[
 [
  which
  np (conc (commonName, plural, [Trait], [04616059]), A, n(traits))
  vp (conc (plural, _, [00137313, 00019448, 02677097]), B, v(are), v(affected))
  pp (conc (_, _), C, prep(by), D)
  np (conc (properName, plural, [QTL], _), X*, n(QTLs))
  ?
 ],
 [
  which
  np (conc (commonName, plural, [Trait], [04616059]), A, n(traits))
  vp (conc (plural, _, [00137313, 00019448, 02677097]), B, v(are), v(affected))
  pp (conc (_, _), C, prep(by), D)
  np (conc (properName, sing, [QTL], _), X*, n(QWsv.DiMo-5H.1)),
  ?
 ],
 [
  which
  np (conc (commonName, plural, [Trait], [04616059]), A, n(traits))
  vp (conc (plural, _, [00137313, 00019448, 02677097]), B, v(are), v(affected))
  pp (conc (_, _), C, prep(by), D)
  np (conc (properName, sing, [QTL], _), X*, n(QWsv.DiMo-5H.2))
  ?
 ],
]
```

In Figure 5, a screenshot of the authoring tool is presented for this last context question. The reformulation (its slot structure) selected by the user in the authoring tool will be run and added to the previous_questions structure, which will be used in the following context questions posed by the user.

previous_questions

```
[
 [
     what
     np (conc (properName, plural, [QTL],
  _), X, n(QTLs))
     vp (conc (plural, _, [00713167,
  02724417, 02458103]), Y, v(are),
  v(related), prep(to))
     np (conc (commonName, sing, [Trait],
  [05033410]), Z, n(frost), n(tolerance),
  T)
     pp (conc (_, _), V, prep(in), W)
     np (conc (commonName, sing,
  [Species], [12142085, 07803545]), W,
  n(wheat))
     ?
 ],
 [
     what
     np (conc (properName, plural, [QTL],
  _), X*, n(QTLs))
     vp (conc (plural, _, [00713167,
  02724417, 02458103]), Y*, v(are),
  v(related), prep(to))
     np (conc (commonName, sing, [Trait],
  [05033410]), Z*, n(frost),
  n(tolerance), T)
     pp (conc (_, _), V, prep(in), W)
     np (conc (commonName, sing,
  [Species], [07803093, 12123244]), W,
  n(barley))
     ?
 ]
]
```

previous_answers

```
[
 [
     np (conc (properName, sing, [QTL],
  _), X1, n(QWin.ipk-6A))
 ],
 [
     np (conc (properName, sing, [QTL],
  _), X2, n(QWsv.DiMo-5H.1)),
     np (conc (properName, sing, [QTL],
  _), X3, n(QWsv.DiMo-5H.2))
 ]
]
```

Figure 4. Slot structures stored in *previous_questions* and *previous_answers* for two context questions

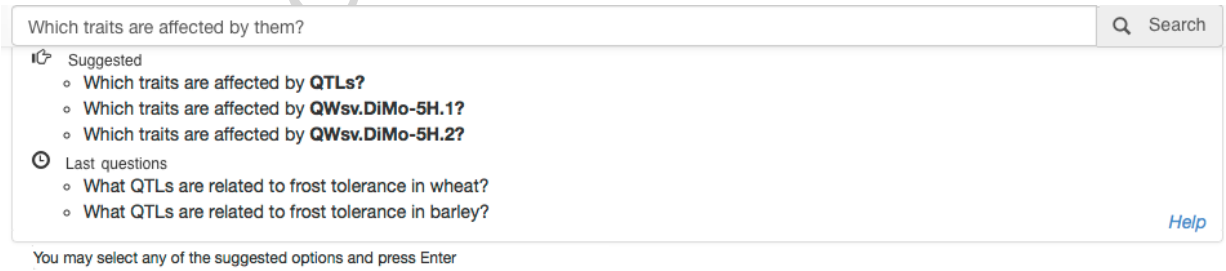


Figure 5. Screenshot of the context question "Which traits are affected by them?" that follows "What QTLs are related to frost tolerance in wheat? What QTLs are related to frost tolerance in barley?"

4. Case study

In this section, we introduce the case study in which we have tested the application of our proposal through the following four subsections: general description of the case study; ontology generation of the CEREALAB database; the analysis of a set of series of context questions processed by our proposal; and some examples of the dashboards generated in the test and presented to the user.

4.1.Description

We illustrate the application of our proposal by means of the following case scenario: a plant breeder enterprise wants to carry out new breeding programs experimenting with the new advances in Genetics. This case scenario will use our framework in Peral et al. (2015), in which the decision maker of the breeding program can easily access to external data about relevant agronomic traits and draw up new molecular protocols to design genetically modified crops in order to increase the productivity of the seed industry. As described in the previous section, our authoring tool to deal with anaphora, ellipsis and context questions is embedded in the GUI module of this framework, in which the user poses the NL questions to query the CEREALAB database (Milc et al., 2011). The CEREALAB database aims to store genotypic and phenotypic data obtained by the CEREALAB project and to integrate them with already existing data sources in order to create a tool for plant breeders and geneticists. The database can help them in unravelling the genetics of economically important phenotypic traits; in identifying and choosing molecular markers associated to key traits; and in choosing the desired parentals for breeding programs. The database is divided into three sub-schemas corresponding to the species of interest: wheat, barley and rice; each sub-schema is then divided into two sub-ontologies, regarding genotypic and phenotypic data, respectively. Although some databases designed to store and manage both phenotypic and genotyping data have been reported, such as AppleBreed (Antofie et al., 2007) or PlantDB (Exner et al., 2008) among others, we have decided to use CEREALAB because those databases are often designed to store the experimental data and the data available are generally restricted to those implemented by the developers/users with no possibility to take advantage of already available information that resides in other data sources. Moreover, CEREALAB is the first database specific for breeding of wheat, barley and rice, fundamental crops for the world agriculture.

The corresponding model for the mentioned scenario, shown in Figure 6, is based on a UML profile for modelling DWs presented in Luján-Mora, Trujillo, & Song (2006). It captures the structure of the initial information to be analyzed. We can see four different dimensions in our model: DNA Sequence, Trait, Species and Effect. Firstly, the DNA Sequence dimension captures the information regarding the QTLs and Genes involved in the different traits shown by the various species of plants. The DNA Sequence dimension is composed by three hierarchy levels, each of them identified by the corresponding scientific code or name given to the element. QTLs and Genes are grouped into their corresponding Chromosomes that represent the highest level of aggregation. Secondly, the Trait dimension captures the traits affected by the presence of the QTLs/Genes. Traits are identified by the code name assigned to them. Some examples can be “Frost resistance” or “Ash content”. Traits also can have a description and can be related to other traits, captured by means of the SeeAlso attribute. Finally, if the trait has been extracted from a data source, it is stored within the DataSource attribute. Thirdly, the Species dimension captures the information about the varieties of plants which has the QTLs/Genes. This dimension contains all the information about each variety, including the Genus, Sub-Family and Family. Each of these levels includes the corresponding identifier of the group that the variety pertains to. In our case, we will only store information about Wheat, Barley and Rice at the highest level of the hierarchy, although additional

information could be added regarding other groups. Fourthly, the Effect dimension captures the effect that a certain QTL/Gene has on a Trait of a Species. The reason to include this separate dimension is because most data warehouse technologies are designed to contain numerical values within the fact. As the effect of the QTLs/Genes on the traits presents a wide variety, from changing colours shown to changing the percentage of certain chemical elements present in the plant, we add this dimension to store this information. Finally, our fact includes a measure that provides an idea of how much evidence there is in terms of the number of studies that support the effect of a QTL/Gene on the trait of a plant species. This information is retrieved from the enterprise internal data. The measure is aggregated with the addition of evidence encountered that a trait is affected by a QTL/Gene.

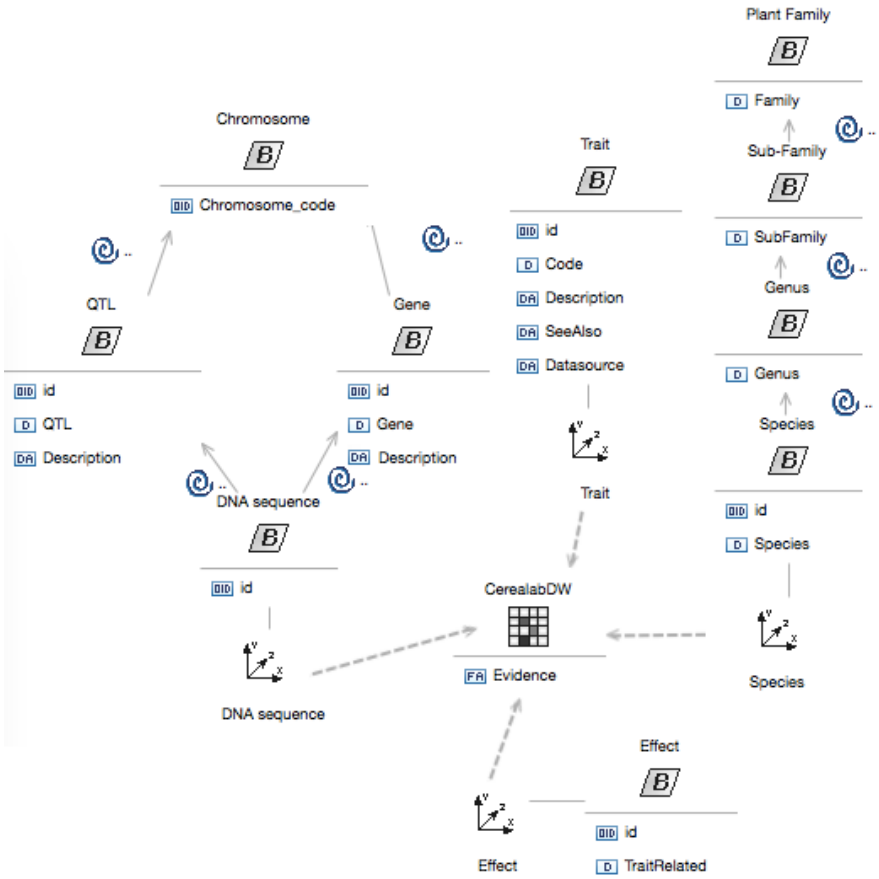


Figure 6. Excerpt of the multidimensional model for CEREALAB scenario

4.2.Ontology generation

In this subsection, the ontology generation for the CEREALAB database (our case study) is illustrated according to the five-steps method proposed in subsection 3.1 whose result is depicted in Figure 7. In the first step, the DNA Sequence dimension is transformed into the concepts of DNA Sequence, Gene, QTL and Chromosome; the Species dimension is transformed into the concepts of Species, Genus, Sub-Family

and Plant Family; in the same way, the concept Trait is created. In the second step, the node CerealabDW is related (using a *has_members* relationship) to the finest grained levels on the three dimensions (DNA Sequence, Species, and Trait). During the third step, all the drill down and roll up relationships are included in the ontology as *has_members* and *is_a_member_of* relationships respectively; for instance, the link *is_a_member_of* is added from Species to Genus and *has_members* is added from Genus to Species. In the fourth step, the designer renames the created relationships and enriches the ontology by including the relationships across different dimensions. We can divide these relationships into two main groups, standard relationships (similar to those used in WordNet) and specific relationships (new ones special for the domain of Genetics and Biotechnology).

With regard to the first group, the relationships included are the following: (1) *synonym* to express that different words represent the same concept (for instance, the terms "durum wheat" and "Triticum durum" are instances of the same Species concept). (2) *has_parts*/*is_a_part_of* to indicate the meronymy relation; for example, a Chromosome consists of several Genes (*has_parts*) whereas a Gene *is_a_part_of* a Chromosome. (3) *has_particulars/is_a_kind_of* to specify hyponymy/hyperonymy relations, as occurs with QTL and DNA Sequence: a QTL is a type of DNA Sequence (*is_a_kind_of*). (4) *has members/is_a_member_of* to designate the membership relation as shown in the nodes related to the name of the Species; for instance, a Genus has several Species (*has_members*) whereas a Species *is_a_member_of* a Genus. Some examples of these relationships with the concrete instances of the Cerealab DW are: the QTLs QWsv.DiMo-5H.1 and QWsv.DiMo-5H.2 *are parts of* the 5H chromosome in barley; the Species *Triticum durum is a member of* the Genus Triticum.

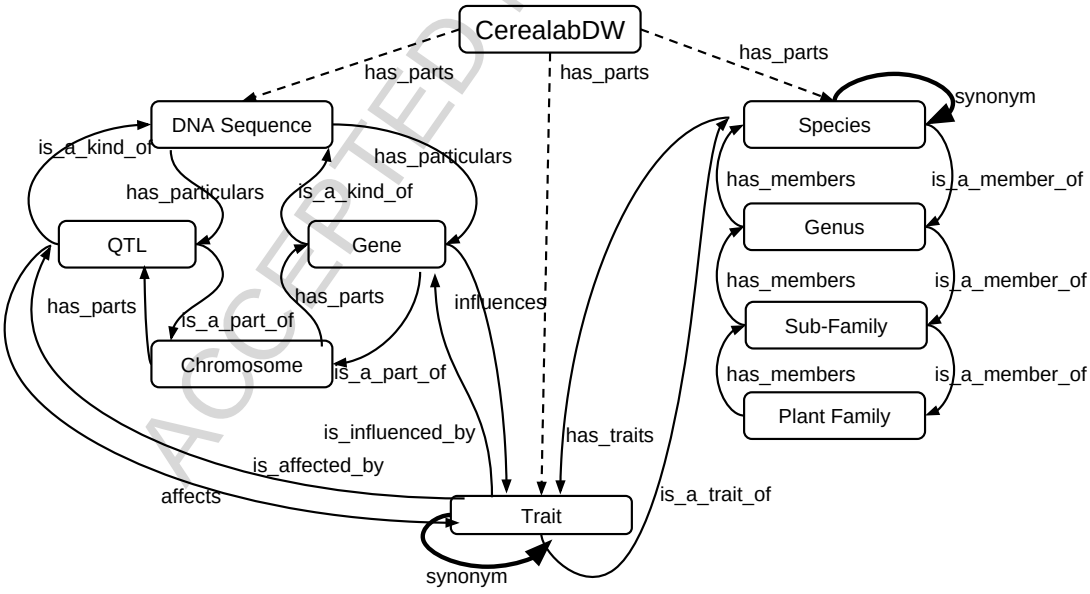


Figure 7. Ontology created for the CEREALAB database

In respect of the second group of specific relationships for the particular domain, they are the following: (1) *has_traits/is_a_trait_of* to denote that each Species has several specific Traits (*has_traits*). (2) *influences/is_influenced_by* to determine that a Gene decisively *influences* on a given Trait whereas

the relationship (3) *affects/is_affected_by* is very similar but more appropriated for the term QTL. Examples of these relationships are: the Species *Hordeum vulgare has the traits* "frost tolerance", "aleurone colour", and "1000 kernel weight" among others; the Genes Dn1, Dn2, Dn3, and Dn4 among others *influence* the "resistance to Russian wheat aphid" Trait in wheat; the QTLs AQGD028, AQGD029, and AQGD030 *affect* the "pericarp colour" Trait in rice.

The created ontology has been formalized using the Web Ontology Language (OWL) following W3C Recommendations (Dean & Schreiber, 2004; Patel-Schneider, Hayes & Horrocks, 2004). We have used Protégé 4 (ontology editing environment) to create the ontologies (http://protege.stanford.edu/).

4.3.Series of context questions

In this subsection, the application of our proposal (see Algorithm 1) to a set of series of questions is shown. The NLI with which it has been implemented is the one developed in Llopis & Ferrández (2012). Moreover, the used anaphora resolution tool is the one that we proposed in Palomar et al. (2001) and Muñoz et al. (2000), which works on the output of a POS tagger, TreeTagger[1] for English and Maco+[2] for Spanish, and on the output of our SUPAR parser (Ferrández et al., 1999). This anaphora resolution tool achieves a precision of 81% in Spanish pronouns, 74% in English, 78% in definite descriptions, all these figures on general purpose corpora. In respect of the implementation of the framework in Peral et al. (2015), it is using the open-source BI platform called Pentaho, which provides the necessary OLAP capabilities by means of the Mondrian OLAP server. The OLAP server was connected to a MySQL Server 5.6 DBMS that stores the data for the analysis. The following series of questions were proposed by two experts in Biotechnology and Agriculture, which had previous knowledge about the CEREALAB database:

a)

1. Tell me the QTLs related to awn colour in wheat.
2. And to glume colour?

---

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (visited on 6th of January, 2015).

[2] http://nlp.lsi.upc.edu/freeling/ (visited on 6th of January, 2015).

3. To semolina one?

b)
1. Does any QTL affect on days to flowering in wheat?
2. In which chromosome are they located?

c)
1. I want to increase grains per spike in barley. Which QTLs are related to this?
2. What other traits are affected by them?

d)
1. I would like to know the QTLs related to rust in wheat
2. Which genes influence the resistance to plagues in wheat?
3. Are there other plagues in wheat?

e)
1. What QTLs are related to the kernel weight in barley?
2. Are there any biotechnology companies that have made Genetic Engineering with these QTLs in barley?
3. Are there currently any transgenic barley varieties in the market?

f)
1. What QTLs are related to frost tolerance and resistance to fusarium in barley?
2. Since there is no common QTL, are there any studies showing these traits are related?

g)
1. Are there currently any transgenic wheat?
2. Is it in the market?
3. What is its price?

Next, we will detail the resolution process of these series of questions using our framework. In series (a), after the partial parsing of the first question ("Tell me the QTLs related to awn colour in wheat"), the following slot structures are obtained (parsedSentence in Algorithm 1):

```
[ vp (conc (_), X, v(tell))
  pron (conc (firstPerson, sing), Y, me)
  np (conc (properName, plural), Z, determiner(the), n(QTLs))
  vp (conc (_), U, v(related), prep(to))
  np (conc (commonName, sing), V, n(awn), n(colour))
  pp (conc (), W, prep(in), A)
  np (conc (commonName, sing), A, n(wheat))
  ?
]
```

The three extracted NP slot structures are the following: "the QTLs", "awn colour", and "wheat". There is not any third-person pronominal anaphor to solve, and there is a definite description ("the QTLs") that cannot be resolved because it is the first NP of the first question, so there are not possible antecedents to which it could refer. In the concept matching with the ontology these slot structures correspond to the ontology node "QTL", an instance of the node "Trait", and an instance of "Species" respectively. There is not ellipsis, and then it is not necessary to reformulate the question. Then, the question is sent to the Distributor/Integrator node (see Figure 1) and the answer is obtained: QRaw.ipk-1A, QRaw.ipk-1D.

For question (a.2): "And to glume colour?". After the parsing, "to glume colour" is detected and tagged as "Trait". No anaphor is detected, but an ellipsis is detected because there is not a verb in the question. It

is solved by comparing the slot structures of the current question with those of the previous one. The missing concepts are "QTL" and "Species". The sentence is completed taking into account the syntactic structure of the previous sentence (i.e. the concept that is not restored from the previous sentence is the one that matches with the one in the present sentence: "awn colour" vs "glume colour"); in this example, the suggested sentence is "Tell me the QTLs related to glume colour in wheat?" The answer QRg.ipk-1D is presented to the user.

In question (a.3): "To semolina one?", the "semolina one" NP is parsed. In the anaphora resolution module, the one anaphora is solved and replaced by its "semolina colour" antecedent, because of the semantic similarity (Trait) between "semolina" and "glume". Then, an ellipsis is found because the sentence does not contain a verb. The missing concepts compared to the previous sentence are "QTL" and "Species". With similar syntactic structure than the previous sentence, the new reformulated sentence is "Tell me the QTLs related to semolina colour in wheat" obtaining the answer: QY.ucw-1B, QY.ucw-6A, QY.ucw-7A, QY.ucw-7B.

Let us analyse the application of our proposal on series (b). In the first question, "Does any QTL affect on days to flowering in wheat?", three NPs slot structures have been identified: "QTL", "days to flowering", and "wheat". There is not any anaphoric expression. The concepts "QTL", an instance of "Trait", and an instance of "Species" have been matched for each NP respectively. No ellipsis has been detected. Finally, the question is sent to the Distributor/Integrator node and returns the answer: QFlt.ipk-3A, QFlt.wak-2D.

In question (b.2): "In which chromosome are they located?", once the parsing has been carried out, the anaphora resolution module detects a plural third person pronoun. The antecedent is searched in the concepts of the current sentence, the previous one, and the answers of the previous question. This is a highly ambiguous anaphor because the list of possible antecedents is quite long. The list of possible solutions is headed by the only plural NP "days", followed by the remaining NPs sorted by probability of certainty (highly influenced by proximity to the anaphor). No ellipsis is detected. Finally, the system generates the following questions, from which the user selected: "In which chromosome are QFlt.ipk-3A located?", "In which chromosome are QFlt.wak-2D located?".

In series (c), the user wants to know if the QTLs that affect a trait (grains per spike) influence other traits (collateral effects), such as the grain weight. In the first question, "I want to increase grains per spike in barley. Which QTLs are related to this?", the singular third person pronoun "this" is found. It is an especially difficult case to solve, since the antecedent is not a NP but a verbal phrase. In this case, the user had to solve it manually through the authoring tool, reformulating the question as "Which QTLs are related to increase grains per spike in barley?". In the next stage, the system cannot match the NP "grains per spike" against any node in the ontology. The NP tagging using WordNet presents as synonym the concept "kernels per spike", which is found in the ontology as an instance of "Trait". The remaining NPs are identified in the ontology. Finally, ellipsis is not found and the following answers are obtained: QKer.pil-1H, QKps.BlKy-1H, QKps.BlKy-3H, QKps.BlKy-4H, QKps.BlKy-5H, QKps.BlKy-6H, QKps.TyVo-2H.

With regard to (c.2) question, "What other traits are affected by them?", after the partial parsing, the pronoun "them" is solved and replaced by its possible antecedents (including the answers of the previous question that consists of several NPs). Later, all the concepts are identified. No ellipsis is detected in the question. To conclude, the system reformulates one question for each single answer of the previous question: "What other traits are affected by QTLs?", "What other traits are affected by QKer.pil-1H?", "What other traits are affected by QKps.BlKy-1H?", "What other traits are affected by QKps.BlKy-3H?", etc.

In series (d.1), “I would like to know the QTLs related to rust in wheat”, three NPs have been detected: “QTLs”, “rust”, and “wheat”. No anaphor is found. After that, all the concepts are matched with a single ontology node except “rust” that matches with three instances of “Trait”, “resistance to leaf rust”, “resistance to stem rust”, and “resistance to stripe rust”. Ellipsis is not detected and, finally, a question for each “Trait” is suggested to the user: “I would like to know the QTLs related to resistance to leaf rust in wheat.”, “I would like to know the QTLs related to resistance to stem rust in wheat.”, “I would like to know the QTLs related to resistance to stripe rust in wheat.”.

In question (d.2), “Which genes influence the resistance to plagues in wheat?”, once the three NPs have been parsed and no anaphor is detected, begins the tagging of concepts. “Genes” and “wheat” are matched whereas “plagues” is not found in the ontology. After that, it is expanded by means of WordNet, locating “group/swarm of insects” as a synonym of “plague”, which indicates that it is a “Trait” of the wheat. Using the relations of WordNet and the ontology concepts, the system determines that the traits “resistance to Russian wheat aphid” and “resistance to hessian fly” refer to plagues because “aphid” and “fly” are insects. Ellipsis is not detected and, finally, a question for each “Trait” is suggested to the user: “Which genes influence the resistance to Russian wheat aphid in wheat?”, “Which genes influence the resistance to hessian fly in wheat?”

Next, question (d.3) “Are there other plagues in wheat?” is resolved like the previous questions. A remarkable issue is that the questions are always classified as the search of concepts marked by the first NP in the question, instead of a question whose answer is just yes/no. Moreover, we should highlight that in our framework (Peral et al., 2015) the user can decide if he/she wishes to consult external data in order to answer the question. In this case, the QA node of the proposed architecture is the responsible to provide the answer.

In the context questions in series (e), the user wants to know information about the competitors (external data) regarding to a specific topic. In the second question, “Are there any biotechnology companies that have made Genetic Engineering with these QTLs in barley?”, after the parsing, the following NPs are detected: “biotechnology companies”, “Genetic Engineering”, “these QTLs”, and “barley”. The anaphora resolution module solves the anaphoric expression “these QTLs” and replaces it by the answer of the previous question (QGwe.HaTR-5H.1, QGwe.HaTR-5H.2, QGwe.HaTR-7H.1, …). At the end, a question for each single answer is reformulated: “Are there any biotechnology companies that have made Genetic Engineering with QGwe.HaTR-5H.1 in barley?”, “Are there any biotechnology companies that have made Genetic Engineering with QGwe.HaTR-5H.2 in barley?”, “Are there any biotechnology companies that have made Genetic Engineering with QGwe.HaTR-7H.1 in barley?”, etc. The system is unable to answer these questions by searching in the CEREALAB database and the user has to decide if he/she wants to consult external data.

In the context questions in series (f), the user is looking for relations between different phenotypic traits. For example, if we want that the barley was resistant to frost, it may occur that it was also resistant to fusarium. In the first question, “What QTLs are related to frost tolerance and resistance to fusarium in barley?”, the NPs “frost tolerance” and “resistance to fusarium” are detected and matched as instances of “Trait” node.

With series (g) the user wants information about the competitors (product prices) in order to determine competitive prices for his products. In these questions, the user is looking for transgenic wheat price. The system is unable to find it in the CEREALAB database and the user has to decide if external information is consulted.

In order to prove the robustness and the usefulness of the approach, the two experts were requested to extend the initial set of questions. Finally, a set of 149 questions (listed in Appendix A) was obtained. Different evaluation measures were used to calculate the proposal performance. Two types of measures were combined: the classic measures used in the iCLEF[3] conferences until 2005 to calculate the accuracy of a system; and measures focused on the user, used from iCLEF 2006. We have used iCLEF measures because this competition focused on user-inclusive perspective applied on search capabilities, simulating the user interaction that we propose in our approach.

In respect of classical measures, we have used the "accuracy" measure (Gonzalo et al., 2006), the fraction of questions in which the user obtained the information searched within a time limit of three minutes (when an user did not find the correct question within the three minutes allowed per question, the question was considered as not resolved similar to the criteria used in iCLEF 2004 and 2005). Here it has only been taken into account if the query results are correct (anaphor and ellipsis resolution has not been evaluated with this measure). It was obtained an accuracy of 81.4% (errors are mainly due to incorrect detection of concepts in the question).

With regard to the user-centred measures, we have used user satisfaction and measures concerning to question reformulations. The user satisfaction has been evaluated with a survey according to the evaluation measures used in iCLEF 2006 (each one valued between 1 and 5): (a) Happy: "Are you satisfied with how you performed the task?" (b) Complete: "Did you find enough, or would you have continued if there had not been a time limit?" (c) Quality: "Compare the illustrations with some given set. Are any of these better than your retrieved results?" The survey results showed a high user satisfaction (with a score of 4.3 out of a maximum of 5 points).

As proposed by the Swedish Institute of Computer Science, SICS, in iCLEF 2009 (Gonzalo et al., 2010), we have used several measures related to question reformulations: (i) correctness (the fraction of questions for which there is at least one right reformulation offered by our system), obtaining a 75.6% (errors were mainly due to fails in anaphora and ellipsis resolution); (ii) average number of

[3] Interactive Cross-Language Retrieval track in Cross-Language Evaluation Forum, CLEF, Editions. http://www.clef-initiative.eu/home (visited on 1st of June, 2015).

reformulations, with 1.4 reformulations for every question (71.1% of the questions were correctly reformulated at the first attempt); (iii) the average time used to answer a question (including the time to do the reformulations), obtaining 35.7 seconds per question.

4.4.Dashboards presented to the user

In this subsection, we are introducing some examples of the dashboards generated in the test and presented to the user. In the top part of Figure 8 and Figure 9, the authoring tool is depicted, which contains:

- The current question posed to the user (e.g. in Figure 8, the question "To semolina one?").
- The suggested reformulation questions after the ellipsis and anaphora resolution.
- The last questions posed by the user.

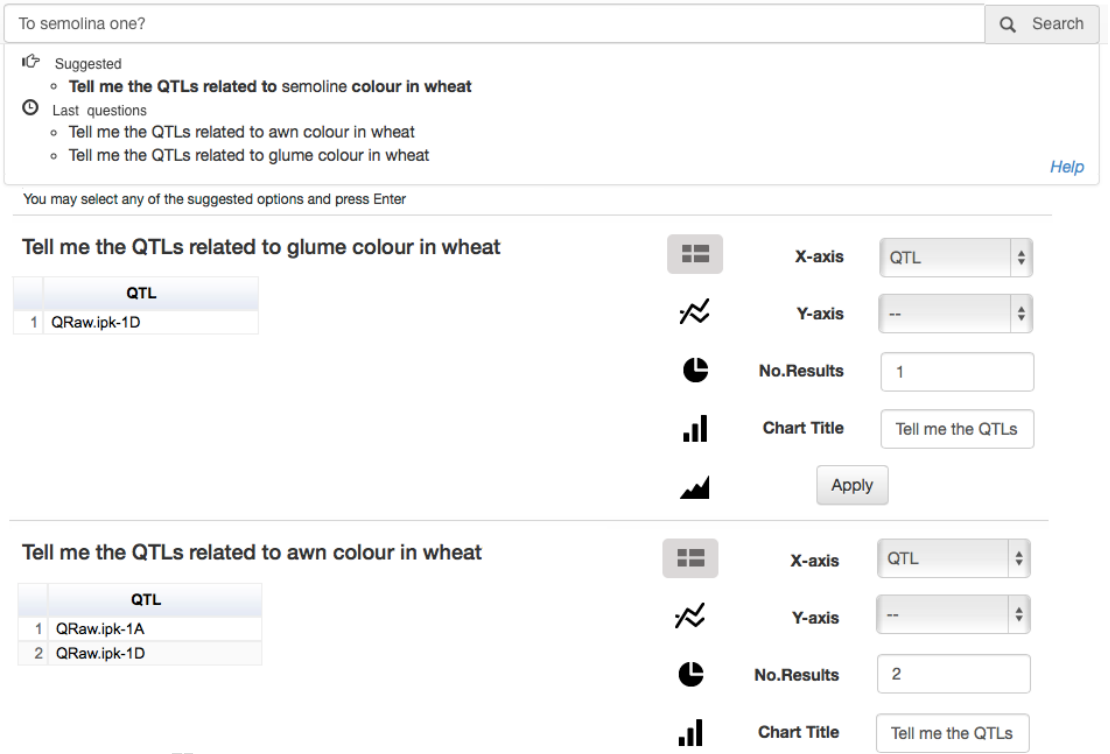


Figure 8. Dashboard for series (a) in subsection 4.3

In the bottom part of these Figures, the dashboards with the solution obtained by the framework for the previous questions in the series are presented. There will be a dashboard for each previous question posed by the user. Each dashboard is titled with the question posed by the user, and is followed by the information extracted from the source nodes (internal data from the database, or external data from the web). On the right hand of this part, additional details about the information extracted are presented.

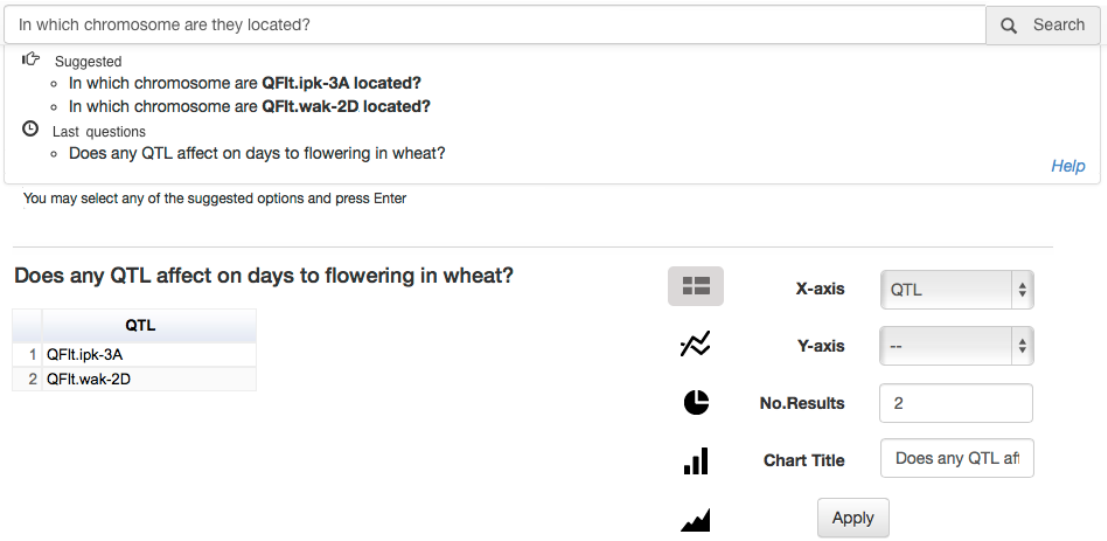


Figure 9. Dashboard for series (b) in subsection 4.3

5. Conclusions and future research

Decision support systems (DSS) support business or organizational decision-making activities. Their main components are the database (or knowledge base), the model (i.e., the decision context and user criteria), and the user interface. In this paper, we face some of the problems in the first and last components of a DSS. Regarding the user interface in a DSS, graphical interfaces present significant limitations to query its knowledge base (e.g. lack of flexibility in query formulation, since the expressivity power is reduced to the user interface design and they are not prepared for arbitrary questioning). Therefore, natural language interfaces (NLI) are expected as the optimal solution. However, especially for non-expert users, a real natural communication is the most difficult to realize effectively.

In this paper, we extend our NLI (Llopis & Ferrández, 2012) that overcomes the problems detected in previous state-of-the-art of NLIs. It improves the interaction between the user and the DSS by means of referencing previous questions or their answers (i.e. it resolves anaphora ambiguity as the pronominal reference in "What traits are affected by them?"), or by eliding parts of the question (i.e. it resolves ellipsis ambiguity as in the question "And to glume colour?" posed after "Tell me the QTLs related to awn colour in wheat"). It provides an integrated solution that handles the context of previous questions and their answers through the resolution of ellipsis and anaphora ambiguity, in order to reach a real natural communication. Moreover, it overcomes one of the main problems of NLIs about the difficulty to adapt the NLI to a new domain, by means of ontologies that are obtained in a semi-automatic way, overcoming the drawback of previous work about its handcrafted generation of the ontologies. Because of the high NL ambiguity of the resolution process (which achieves not enough precision), our proposal is presented as an authoring tool that helps the user to query efficiently in NL.

With regard to the second component of a DSS, the database or knowledge base, we have embedded our NLI in our previous proposal in Peral et al. (2015) that allows the integration of internal and external, structured and unstructured information. In this way, our proposal can interface with databases, data warehouses, question answering, information retrieval and information extraction systems. This tool has

been tested on a DSS case scenario about Biotechnology and Agriculture, where different users query the CEREALAB database through a set of context questions, reaching a high user satisfaction in the interaction with the DSS, facilitating the decision making process.

As future projects, the authors plan to check the adaptability of our proposal on other domains as well as the one studied in this paper. For instance, domains related with the NLI interaction with unstructured information from Social Networks, such as Twitter, which require a pre-processing phase because of the text informality. Furthermore, in Twitter domain, the short length of the tweets and the lack of context will probably affect to the performance of the anaphora and ellipsis resolution modules. Moreover, we will prove the modularity feature of our architecture by means of the use of other state-of-the-art natural language processing tools, such as anaphora resolution (e.g. the Berkeley coreference resolution system in Durrett & Klein, 2013), ellipsis resolution or word sense disambiguation.

6. Acknowledgements

This paper has been partially supported by the MESOLAP (TIN2010-14860), GEODAS-BI (TIN2012-37493-C03-03), LEGOLANG-UAGE (TIN2012-31224) and DIIM2.0 (PROMETEOII/2014/001) projects from the Spanish Ministry of Education and Competitivity. Alejandro Maté is funded by the Generalitat Valenciana under an ACIF grant (ACIF/2010/298).

7. References

Aleksejeva, I. (2014). EU experts' attitude towards use of GMO in food and feed and other industries. *Procedia - Social and Behavioral Sciences*, 110, pp. 494-501.

Altman, R., Bergman, C., Blake, J., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W., Hirschman, L., Jensen, LJ., Krallinger, M., Mons, B., O'Donoghue, SI., Peitsch, MC., Rebholz-Schuhmann, D., Shatkay, H., Valencia, A. (2008). Text mining for biology - the way forward: opinions from leading scientists. *Genome Biology*, 9 Suppl 2:S7, pp. 1-15.

Antofie, A., Lateur, M., Oger, R., Patocchi, A., Durel, C., Van de Weg, W. (2007). A new versatile database created for geneticists and breeders to link molecular and phenotypic data in perennial crops: the AppleBreed DataBase. *Bioinformatics*, 23(7), pp. 882-891.

Barbosa, J.J.G., Rangel, R.A.P., Cruz, I.C., Fraire, H.J., Aguilar, S., Pérez, J. (2006). Issues in translating from natural language to SQL in a domain-independent natural language interface to databases. In Proceedings of the MICAI 2006, LNAI 4293, pp. 922-931.

Bellot, P., Crestan, E., El-Bèze, M., Gillard, L., de Loupy, C. (2002). Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC 11 Question Answering Track. In Proceedings of The Eleventh Text REtrieval Conference (TREC 2002).

Carbonell, J. G. (1983). Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces. In Proceedings of the 21st annual meeting on Association for Computational Linguistics (ACL '83), pp. 164-168.

Chai, J. Y., Jin, R. (2004). Discourse Structure for Context Question Answering. In Proceedings of the HLT-NAACL 2004: Workshop on Pragmatics of Question Answering, pp. 23-30.

Cimiano, P., Haase, P., Heizmann, J., Mantel, M., Studer, R. (2008). Towards portable natural language interfaces to knowledge bases - the case of the ORAKEL system. *Data & Knowledge Engineering*, 65(2), pp. 325-354.

Cimiano, P., Minock, M. Natural language interfaces: What is the problem - A data-driven quantitative analysis. In Proceedings of the NLDB 2009, pp. 192-206.

Clegg, AB., Shepherd, AJ. (2007). Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics*, 8(24), pp. 1-17.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Wilks, Y. (2000). Experience of using GATE for NLP R/D. In Proceedings of the Workshop on Using Toolsets References 2000 and Architectures To Build NLP Systems at COLING-2000, pp. 1-8. http://gate.ac.uk/.

Dean, M., Schreiber, G. (2004). OWL Web Ontology Language Reference. W3C Recommendation, http://www.w3.org/TR/owl-ref/ (visited on 6th of January, 2015).

Díaz de Ilarraza, A., Rodríguez Hontoria, H., Verdejo, F. (1990). A Mechanism for ellipsis resolution in dialogued systems. In Proceedings of the 13th conference on Computational linguistics (COLING '90), pp. 452-454.

Distelhorst, G., Srivastava, V., Rosse C, Brinkley, J. (2003). A prototype natural language interface to a large complex knowledge base, the Foundational Model of Anatomy. In Proceedings of the AMIA Annual Symposium, pp. 200-204.

Durrett, G., Klein, D. (2013). Easy victories and uphill battles in coreference resolution. In Proceedings of the EMNLP, pp. 1971-1982.

Elhai, J., Taton, A., Massar, J.P., Myers, J.K., Travers, M., Casey, J., Slupesky, M., Shrager, J. (2009). BioBIKE: a web-based, programmable, integrated biological knowledge base. *Nucleic Acids Res*, 37:W28-W32.

ETC (Action Group on Erosion, Technology and Concentration) (2008).Who owns nature? Corporate Power and the Final Frontier in the Commodification of Life. Communiqué, Issue N° 100. ETC Group, 52 pp.

Exner, V., Hirsch-Hoffmann, M., Gruissem, W., Hennig, L. (2008). PlantDB - a versatile database for managing plant research. *Plant Methods*, 4:1, doi:10.1186/1746-4811-4-1.

Falconer, D.S., Mackay, T.F.C., (1996). Introduction to Quantitative Genetics, Ed 4.,Green, Longmans.

Ferrández, A., Palomar, M., Moreno, L. (1999). An Empirical Approach to Spanish Anaphora Resolution. *Machine Translation*, 14 (3/4), pp. 191-216.

Ferrández, S., Toral, A., Ferrández, O., Ferrández, A., & Muñoz, R. (2009). Exploiting Wikipedia and EuroWordNet to Solve Cross–Lingual Question Answering. Information Sciences, 179(20), pp. 3473-3488.

Fukumoto, J., Kato, T., Masui, F. (2003). Question Answering Challenge (QAC-1): An Evaluation of Question Answering Tasks at the NTCIR Workshop 3. In New Directions in Question Answering, pp. 122-133.

Fukumoto, J., Kato, T., Masui, F. (2004). An evaluation of question answering challenge (QAC-1) at the NTCIR workshop 3. In Proceedings of the ACM SIGIR Forum, Vol. 38, No. 1, pp. 25-28.

Fukumoto, J., Kato, T., Masui, F., & Mori, T. (2007). An overview of the 4th question answering challenge (QAC-4) at NTCIR workshop 6. In Proceedings of the Sixth NTCIR Workshop Meeting, pp. 433-440.

Giampiccolo, D., Forner, P., Herrera, J., Peñas, A., Ayache, C., Forascu, C., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R. (2007). Overview of the CLEF 2007 Multilingual Question Answering Track. In Proceedings of the CLEF 2007 Workshop, pp. 200-236.

Goldsmith, E.J, Mendiratta, S., Akella, R., Dahlgren, K. (2009). Natural language query in the biochemistry and molecular biology domains based on cognition search. In Proceedings of the AMIA Joint Summits on Translational Science, pp.32-37.

Gonzalo, J., Clough, P., Vallin, A. (2006). Overview of the CLEF 2005 Interactive Track. In Proceedings of the CLEF 2005 Workshop. LNCS 4022, pp. 251-262.

Gonzalo,J., Peinado, V., Clough, P., Karlgren, J. (2010). Overview of iCLEF 2009: Exploring Search Behaviour in a Multilingual Folksonomy Environment. In Proceedings of the CLEF 2009 Workshop. LNCS 6242, pp. 13-20.

Grosz, B.J., Joshi, A.K., Weinstein, S. (1995). Centering: a framework for modeling the local coherence of discourse, *Computational Linguistics*, 21(2), pp. 203-225.

Grosz, B.J., Sidner, C. (1986). Attention, intention, and the structure of discourse. Computational Linguistics, 12 (3), pp.175-204.

Harabagiu, S., Moldovan, D., Pasca, M., Surdeanu, M., Mihalcea, R., Girju, R., Rus, V., Lactusu, F., Morarescu, P., Bunescu, R. (2001). Answering Complex, List and Context Questions with LCC's Question-Answering Server. In Proceedings of the Tenth Text REtrieval Conference (TREC-10), pp. 355-451.

Hobbs, J.R. (1985). On the coherence and structure of discourse. Technical report no. CSLI-85-37, Center for the Study of Language and Information, Stanford University.

ISAAA. (International Service for the Acquisition of Agri-biotech Applications). Pocket K Nº 16. (2012). Global status of commercialized biotech/GM crops. Available at http://www. isaaa.org/resources/publications/pocketk/16/default.asp.

Jamil, H.M. (2012). A natural language interface plug-in for cooperative query answering in biological databases. *BMC Genomics*, 13(Suppl 3):S4, pp. 1-12.

Jensen, L.J., Saric, J., Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2), pp. 119-129.

Kamp, H., Reyle, U. (1993). From Discourse to Logic. Kluwer, Dordrecht.

Kato, T., Fukumoto, J., Masui, F. (2004). Question Answering Challenge for Information Access Dialogue – Overview of NTCIR4 QAC2 Subtask 3. In Proceedings of the NTCIR-4 Workshop Meeting.

Kato, T., Fukumoto, J., Masui, F. (2005). An overview of NTCIR-5 QAC3. In Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access.

Kearsey, M.J., (1998). The principles of QTL analysis (a minimal mathematics approach). *Journal of Experimental Botany*, 49 (327), pp. 1619-1623.

Lamel, L., Rosset, S., Ayache, C., Mostefa, D., Turmo, J., Comas, P. (2008). Question Answering on Speech Transcriptions: the QAST evaluation in CLEF. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), pp. 28-30.

Laukaitis, A., Vasilecas, O. (2007). Natural language as programming paradigm in data exploration domain. *Information Technology and Control*, 36, pp. 30-36.

Le Buanec, B. (2008). Evaluation of the seed industry during the past 40 years. *SEED News. Magazine*. Main subject of July/August — XII nº 4.

Li, Y., Yang, H., Jagadish, H.V. (2005). NaLIX: an interactive natural language interface for querying XML. In Proceedings of the SIGMOD 2005, pp. 900-902.

Li, Y., Chaudhuri, I., Yang, H., Singh, S., Jagadish, H.V. (2007). Enabling domain awareness for a generic natural language interface. In Proceedings of the Association for the Advancement of the Artificial Intelligence (AAAI), pp. 833-838.

Llopis, M., Ferrández, A. (2012). How to make a natural language interface to query databases accessible to everyone: an example. *Computational Standard & Interfaces*, 35, pp. 470-481.

Luján-Mora, S., Trujillo, J., Song, I. (2006). A UML profile for multidimensional modeling in data warehouses. *Data and Knowledge Engineering*, 59 (3), pp. 725-769.

Lynch, M., Walsh, B., (1998). Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc., Sunderland, MA.

Mann, W.C., Thompson, S.A. (1987). Rhetorical structure theory: a theory of text organization. Technical report no. ISI/RS-87-190, Information Sciences Institute, University of Southern California.

Matos, S., Arrais, J.P., Maia-Rodrigues, J., Oliveira, J.L. (2010). Concept-based question expansion for retrieving gene related publications from MEDLINE. *BMC Bioinformatics*, 11(212), pp. 1-9.

Matsuda, M., Fukumoto, J. (2005). Answering Questions of IAD Task using Reference Resolution of Follow-up Questions. In Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access.

Meijerink, G.W., Danse, M.G. (2009). Riding the Wave: High Prices, Big Business? The Role of Multinationals in the International Grain Markets. Report 2009-031. Project code 40789. LEI Wageningen UR, The Hague, The Netherland. 87 pp.

Milc, J., Sala, A., Bergamaschi, S., Pecchioni, N. (2011). A genotypic and phenotypic information source for marker-assisted selection of cereals: the CEREALAB database. *Database*.

Miles, C., Wayne, M. (2008). Quantitative Trait Locus (QTL) analysis. *Nature Education*, 1(1), pp. 208.

Muñoz, R.; Palomar, M.; Ferrández, A. (2000). Processing of Spanish Definite Description. *Lecture Notes in Computer Science*, 1793, pp. 526 - 537.

Palomar, M., Ferrández, A., Moreno, L., Martínez-Barco, P., Peral, J., Saiz-Noeda, M., Muñoz, R. (2001). An Algorithm for Anaphora Resolution in Spanish Text. *Computational Linguistics*, 27(4), pp. 545-567

Patel-Schneider, P.F., Hayes, P., Horrocks, I. (2004). OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation, http://www.w3.org/TR/2004/REC-owl-semantics-20040210/ (visited on 6th of January, 2015).

Peral, J., Ferrández, A., De Gregorio, E., Trujillo, J. Maté, A., Ferrández, L.J. (2015). Enrichment of the phenotypic and genotypic DataWarehouse analysis using Question Answering systems to facilitate the decision making process in cereal breeding programs. *Ecological Informatics*, 26, pp. 203-216.

Popescu, A.M., Etzioni, O., Kautz, H.A. (2003). Towards a theory of natural language interfaces to databases. In Proceedings of the 8th international conference on Intelligent user interfaces (IUI'03), pp. 149-157.

Rótolo, G.C., Francis, C., Craviotto, R.M., Viglia, S., Pereyra, A., Ulgiati, S. (2014). Time to re-think the GMO revolution in agriculture. *Ecological Informatics*. In Press.

Santoso, H., Haw, S., Abdul-Mehdi, Z.T. (2010). Ontology extraction from relational database: concept hierarchy as background knowledge. *Knowledge Based Systems*, 24 (3), pp. 457-464.

Stratica, N., Kosseim, L., Desai, B.C. (2005). Using semantic templates for a natural language interface to the CINDI virtual library. *Data & Knowledge Engineering*, 55, pp. 4-19.

Sun, M., Chai, J. Y. (2007). Discourse Processing for Context Question Answering Based on Linguistic Knowledge. *Knowledge-Based Systems*, 20(6), Special Issue on Intelligent User Interfaces, pp. 511-526.

Tomioka, S. (2008). A step-by-step guide to ellipsis resolution. In Kyle Johnson (ed), Topics in Ellipsis, Cambridge University Press, pp. 210-228.

Vieira, R., Poesio, M. (2000). An Empirically Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4), pp. 539-593.

Voorhees, E. M. (2001). Overview of the TREC 2001 Question Answering Track. In Proceedings of the Tenth Text REtrieval Conference (TREC-10), pp. 42-51.

Voorhees, E. M. (2005). Overview of TREC 2004 Question Answering Track. In Proceedings of the 13th Text REtrieval Conference (TREC 2004), pp. 52-62.

Williams, S. (2000). Anaphoric reference and ellipsis resolution in a telephone-based spoken language system for accessing email. In Botley, Simon Philip and Tony McEnery (eds.), Corpus-based and Computational Approaches to Discourse Anaphora, pp. 171-188.

World Health Organization (2002). Foods derived from modern technology: 20 questions on genetically modified foods. Available from: http://www.who.int/foodsafety/publications/biotech/20questions/en/index.php.

Zheng, J. Chapman, W. W., Crowley, R. S., Savova, G. K., (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 44(6), pp. 1113-1122.

Appendix A

In this appendix, we are listing the set of questions used in the evaluation of our proposal:

a.1. Tell me the QTLs related to awn color in wheat.
a.2. And to glume color?
a.3. To semolina one?

b.1. Does any QTL affect on days to flowering in wheat?
b.2. In which chromosome are they located?

c.1. I want to increase grains per spike in barley. Which QTLs are related to this?
c.2. What other traits are affected by them?

d.1. I would like to know the QTLs related to rust in wheat.
d.2. Which genes influence the resistance to plagues in wheat?
d.3. Are there other plagues in wheat?

e.1. What QTLs are related to the kernel weight in barley?
e.2. Are there any biotechnology companies that have made Genetic Engineering with these QTLs in barley?
e.3. Are there currently any transgenic barley varieties in the market?

f.1. What QTLs are related to frost tolerance and resistance to Fusarium in barley?
f.2. Since there is no common QTL, are there any studies showing these traits are related?

g.1. Are there currently any transgenic wheat?
g.2. Is it in the market?
g.3. What is its price?

h.1. What QTLs are related to awn type in barley?
h.2. And to lemma color?
h.3. To aleurone one?

i.1. Does any QTL affect to protein content in barley?
i.2. In which chromosome are they located?

j.1. I want to increase resistance to common bunt in wheat. Which QTLs are related to this?
j.2. What other traits are affected by them?

k.1. I would like to know the QTLs related to biomolecules content in barley (proteins/lipids/beta glucan).
k.2. Which genes influence the resistance to Fusarium in barley?
k.3. Are there other plagues in barley?

l.1. What QTLs are related to the grain yield in wheat?
l.2. Are there any biotechnology companies that have made Genetic Engineering with these QTLs in wheat?
l.3. Are there currently any transgenic wheat varieties in the market?

m.1. What QTLs are related to grain yield and plant height in wheat?
m.2. Since there is no common QTL, are there any studies showing these traits are related?

n.1. Are there currently any transgenic barley?
n.2. Is it in the market?
n.3. What is its price?

o.1. Tell me the QTLs related to the number of kernels per spike in wheat.
o.2. And to resistance to BYDV?
o.3. To Septoria tritici?

p.1. Does any QTL affect on deoxynivalenol accumulation in wheat?
p.2. In which chromosome are they located?

q.1. I want to increase hull cover in barley. Which QTLs are related to this?
q.2. What other traits are affected by them?

r.1. I would like to know the QTLs related to tan spots in wheat.
r.2. Which genes influence the resistance to Russian wheat aphid in wheat?
r.3. Are there other plagues in wheat?

s.1. What QTLs are related to the friability in barley?
s.2. Are there any biotechnology companies that have made Genetic Engineering with these QTLs in barley?

t.1. What QTLs are related to resistance to Schizaphis and resistance to Powdery mildew in barley?
t.2. If there is no common QTL, are there any studies showing these traits are related?

u.1. Tell me the QTLs related to P/L ratio in wheat.
u.2. And to resistance to Powdery mildew?
u.3. To stem rust one?

v.1. Does any QTL affect to resistance to SBWMV in wheat?
v.2. In which chromosome are they located?

w.1. I want to decrease leaf rust seedling in barley. Which QTLs are related to this?
w.2. What other traits are affected by them?

x.1. I would like to know the QTLs related to aspect of kernels in wheat.
x.2. Which genes influence the resistance to black point in wheat?
x.3. Are there other fungal infections in wheat? Which QTLs are involved in the resistance?

y.1. What QTLs are related to the heading date in barley?

y.2. Are there any biotechnology companies that have made Genetic Engineering with these QTLs in barley?

z.1. What QTLs are related to spike morphology and stripe rust in barley?
z.2. If there is no common QTL, are there any studies showing these traits are related?

aa.1. If there is some transgenic variety of rice, what traits are improved?
aa.2. Are there novel QTLs in the literature which are related to a certain trait in rice?
ab.1. Are there novel genes in the literature which are related to a certain trait in rice?
ab.2. In wheat?

ac.1. What QTLs are related to flowering in rice?
ac.2. And to panicle length?
ac.3. To brown rice one?

ad.1. Does any QTL affect on days to flowering in rice?
ad.2. In which chromosome are they located?

ae.1. I want to increase grain yield in rice. Which QTLs are related to this?
ae.2. What other traits are affected by them?

af.1. I would like to know the QTLs related to amylose content in rice.
af.2. Which genes influence the brown rice length in rice?
af.3. Are there other brown rice dimensions which can be affected?

ag.1. What QTLs are related to the plant height in rice?
ag.2. Are there any biotechnology companies that have made Genetic Engineering with these QTLs in rice?
ag.3. Are there currently any transgenic rice varieties in the market?

ah.1. What QTLs are related to panicle length and brown rice length in rice?
ah.2. Are there any studies showing these traits are related?

ai.1. What QTLs are related to hectolitric weight in barley?
ai.2. And to lipid content?
ai.3. To grain nitrogen one?

aj.1. Does any QTL affect on net blotch in barley?
aj.2. In which chromosome are they located?

ak.1. I want to decrease spikelet sterility in rice. Tell me the QTLs related to this trait.
ak.2. What other traits are affected by them?

al.1. I would like to know the QTLs related to spike row number in barley
al.2. Which genes influence the resistance to Diuraphis in barley?
al.3. Are there other plagues in barley?

am.1. What QTLs are related to the kernel texture in wheat?
am.2. Are there any companies working with these QTLs in wheat?

an.1. What QTLs are related to aromaticity and spikelet sterility in rice?
an.2. Which studies showing these traits are related?

ao.1. What QTLs are related to rachilla hair length in barley?
ao.2. And to awn type?
ao.3. To its roughness?

ap.1. Does any QTL affect to leaf scald in barley?
ap.2. In which chromosome are they located?

aq.1. I want to modify glume color in wheat. Which QTLs are related to the glume?
aq.2. What other traits are affected by them?

ar.1. I would like to know the QTLs related to spot blotch in barley.
ar.2. Which genes influence the resistance to Schizaphis in barley?
ar.3. Which other insects affecting barley?
ar.4. Which QTLs are involved in the resistance?

as.1. What QTLs are related to semolina color and kernel color in wheat?
as.2. Which studies showing these traits are related?

at.1. Show me the QTLs of interest for grain yield in wheat.
at.2. Show me the QTLs which are related simultaneously to resistance to Fusarium and to Schizaphis in barley.

au.1. Are there any trait related to awn in barley?
au.2. In wheat?

av.1. Are there any genes responsible of pericarp color in rice?
av.2. Of grain yield?
av.3. In wheat?
av.4. In barley?

aw.1. What QTLs are related to lodging in rice?
aw.2. And to flowering time?
aw.3. To maturity one?

ax.1. Are there more than 1 gene responsible of the grain yield in wheat?
ax.2. Which of these genes are involved in the grain yield in barley?
ax.3. In rice?

ay.1. Does any QTL affect on aromaticity in rice?
ay.2. In which chromosome are they located?

az.1. I want to obtain a Fusarium-resistant variety of wheat. What genes are suitable/appropriate/useful for this work?

ba.1. What attributes are dependent on gene Sh3 in Hordeum?
ba.2. Which feature depends on gene blx1 in Hordeum?

bb.1. I would like to know the QTLs related to panicle blast in rice.
bb.2. Which genes are related to the leaf blast in rice?

bc.1. What QTLs are related to brown rice shape in rice?
bc.2. And to brown rice length?
bc.3. To the plant one ?

bd.1. Tell me the QTLs linked to the awn color in wheat.
bd.2. Which are the relevant genes in the frost tolerance in wheat?

be.1. 1. Does any QTL affect to grain weight in rice?
be.2. In which chromosome are they located?

bf.1. Search for genes associated to grain nitrogen in wheat.
bf.2. In barley.

bg.1. I want to avoid lodging in rice. Which QTLs are related to lodging?
bg.2. What other traits are affected by them?

bh.1. What genes are underlying the lodging event in wheat?
bh.2. What ecological process is underlined by the gene CHL7 in rice?

bi.1. I would like to know the QTLs related to the development of rice plant.

bi.2. Which genes influence the pericarp color in rice?
bi.3. Are those genes related to the color of other parts?

bj.1. What trait is influenced by gene DP1 in rice?
bj.2. What biological feature is affected by QTL C1AS16 in rice?