Information extraction from bibliography for Marker Assisted Selection in wheat

Claire Nédellec, Robert Bossy, Dialekti Valsamou, Marion Ranoux, Wiktoria Golik, Pierre Sourdille

Information Extraction from Bibliography for Marker-Assisted Selection in Wheat

Claire Nédellec[1], Robert Bossy[1], Dialekti Valsamou[1], Marion Ranoux[2], Wiktoria Golik[1],
Pierre Sourdille[2]

[1] INRA, unité UR1077 MIG (Mathématique, Informatique et Génome),
Domaine de Vilvert, 78 352 Jouy-en-Josas
[2] INRA, UMR1095 GDEC (Génétique, Diversité, Ecophysiologie des Céréales),
Domaine de Crouël, 5 chemin de Beaulieu, 63 039 Clermont-Ferrand cedex

[1] {prénom.nom}@jouy.inra.fr
[2] {prénom.nom}@clermont.inra.fr

Abstract: Improvement of most animal and plant species of agronomical interest in the near future has become an international stake because of the increasing demand for feeding a growing world population and to mitigate the reduction of the industrial resources. The recent advent of genomic tools contributed to improve the discovery of linkage between molecular markers and genes that are involved in the control of traits of agronomical interest such as grain number or disease resistance. This information is mostly published as scientific papers but rarely available in databases. Here, we present a method aiming at automatically extract this information from the scientific literature and relying on a knowledge model of the target information and on the *WheatPhenotype* ontology that we developed for this purpose. The information extraction results were evaluated and integrated into the on-line semantic search engine *AlvisIR WheatMarker*.

Keywords: information extraction, corpus annotation, natural language processing, ontology building, biology, genetics.

1 Introduction

A large amount of work has been done in information extraction (IE) from scientific literature in biology during the past decade. Most of this research has been applied to the extraction of genetic regulations in the molecular biology field, such as protein and gene interactions. It has been popularized by shared tasks (LLL [1], BioCreative [2], BioNLP-ST [3]. Nowadays, the extraction of organism trait and phenotype mentions from papers encounters a growing interest [4,5]. This knowledge is critical in many domains notably agriculture and health and it is rarely available in databases. The phenotypes of plant varieties of agronomical interest are described in scientific papers with the genetic information used for the variety selection. Compared to genetic regulations, the extraction of this knowledge is challenging in IE. In the domain of wheat selection assisted by molecular markers, the knowledge to be extracted and formalized belongs to various fields, *e.g.* genetics, physiology, plant environment, food processing. Its representation involves several n-ary relations and entities that are complex to identify in the texts. The terms that denote traits and phenotypes are very diverse and difficult to predict.

The main approach in relation extraction (RE) for biology uses supervised machine learning trained with reference annotated corpora. The annotation follows a schema that defines the type of relations and entities to be extracted. In this paper we describe how we formalized the knowledge of marker-assisted selection in wheat into a text annotation schema that is appropriate both for the annotation of the reference corpus by domain experts and for the automatic extraction and representation of the knowledge (section 3). Common methods for entity prediction include dictionary matching, supervised machine learning and term analysis. This paper presents a multi-strategy named entity recognition (NER) method that takes into account the diversity of the entity naming and the availability of nomenclatures (section 4). The lack of a controlled vocabulary on wheat phenotypes and traits led us to build a domain specific ontology. The results of the NER methods were evaluated with the reference corpus and used in a bibliographical semantic search engine (section 5).

2 Wheat selection assisted by genetic marker

Improvement of most animal and plant species of agronomical interest in the near future has become an international stake because of the increasing demand for feeding a growing world population and to mitigate the reduction of the industrial resources especially the oil. The new environmental constraints such as the reduction of inputs (water, fertilizers and pesticides) and of acreages involve the development of new breeding schemes that must be shorter and more powerful. This increase needs a significant improvement of the agronomical potential of the species through breeding. This is especially true for wheat, the most widely grown crop worldwide.

Until now, the conventional selection methods lead to maintain the yields just covering the current consumption. The recent advent of genomic tools contributed to improve linkage between molecular markers and genes of agronomical interest. This information must now be integrated in breeding programs and the aim is to move from genetic toward genomic selection. A large number of varieties and molecular markers have been developed these last ten years for the bread wheat (see [6] for a review). However, the most useful information has to be extracted from thousands of scientific articles among which only a few are relevant. In addition, within each interesting paper, only a small part deal with the linkage itself, they indicate the name of the closest marker, the gene itself and the protocol that is useful to reveal the appropriate molecular signal that can be used for marker-assisted selection (MAS). Much more than retrieving relevant papers, breeders need to access the information in a structured form.

Our information extraction goal is the extraction of relationships between entities that are molecular markers, genes, traits, phenotypes and varieties from published papers. Traits are defined as observable characters such as the resistance to a given disease. The phenotypes are the values of the traits, *e.g.* the resistance or the susceptibility to a disease. The alleles of the genes are the different versions of the genes leading to the genotype of the individual. They control the phenotypes. An allele is generally attributed to a molecular marker. The marker discriminates the different alleles of a same gene with the polymorphism of the DNA sequence. The molecular markers are used to select the varieties with a phenotype of agronomic interest. The linkage between molecular markers and genes we focused on are related to four main subjects with high economic impact, (1) biotic stress: resistance to diseases (*e.g.* rust, fusarium, septoria), resistance to pest (*e.g.* greenbug, Cecidomyides, Hessian fly); (2) abiotic stress (*e.g.* drought, soil salinity, temperature, lodging), (3) plant development (*e.g.* vernalization, flowering) and (4) bread quality (*e.g.* grain hardness, protein content).

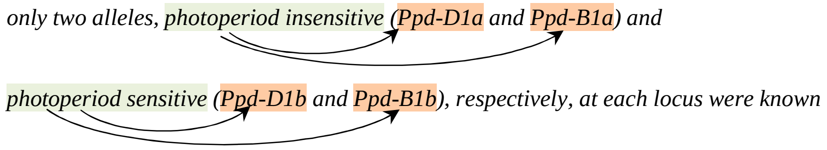
A given knowledge may be expressed in the text of the papers in different ways as shown in example 1. In example 2, we can see that many entities and relations may occur in a single sentence. These features make the information extraction task difficult.

Example 1. The phenotype resistance to leaf rust diseases that is controlled by the gene *Lr34* is expressed by the two very different clauses:

a. *the gene Lr34 confers resistance to leaf rust* [..]
b. [..] *lines missing Lr34 allele are susceptible*.

In clause a., the gene *Lr34* is explicitly designated as controlling the resistance phenotype, whilst clause b. states the same fact in an indirect way: the genotype where the gene *Lr34* has been knockout makes the wheat variety susceptible to the disease. This means that the variety needs the gene *Lr34* to be resistant.

Example 2. [PMID 20002313]



only two alleles, photoperiod insensitive (Ppd-D1a and Ppd-B1a) and

photoperiod sensitive (Ppd-D1b and Ppd-B1b), respectively, at each locus were known

The four allele entities (*Ppd-D1a*, *Ppd-D1b*, *Ppd-B1a*, *Ppd-B1b*) and the two phenotype entities (*photoperiod insensitive* and *photoperiod insensitive)* are the argument of four instances of the *allele_expresses_phenotype* relationships.

Despite this complexity, the recent progress of RE in molecular biology as evaluated in shared tasks open up possibilities of large scale extraction of complex events in the wheat MAS domain.

3 Knowledge model and annotated corpus

The source of information on the linkage between molecular markers and genes in wheat is diverse. We identified 1,229 scientific journals that published relevant papers. These references were obtained by querying Web of Science (WoS) with the *wheat*, *marker* and *gene* keywords. It yielded 3,170 references to scientific papers. Among the retrieved references, we selected 125 relevant journals according to their availability, their impact, their scope, their geographical area and the frequency of relevant publication. A corpus of 2,097 full-texts (*WheatMAS* corpus) was then obtained from the journal publishers that concentrate the target knowledge.

With the breeder experts, we built the MAS knowledge model for the representation of the relevant information of the text. The knowledge model contains 8 entity types and 14 n-ary relationships (10 binary relationships and 4 ternary relationships) that are shown in Figures 1a and 1b. The main entity types are marker, allele and gene, trait and phenotype and variety. *Type* represents the method used to identify the marker, *e.g.* AFLP, microsatellite, which is useful for the evaluation of its quality.

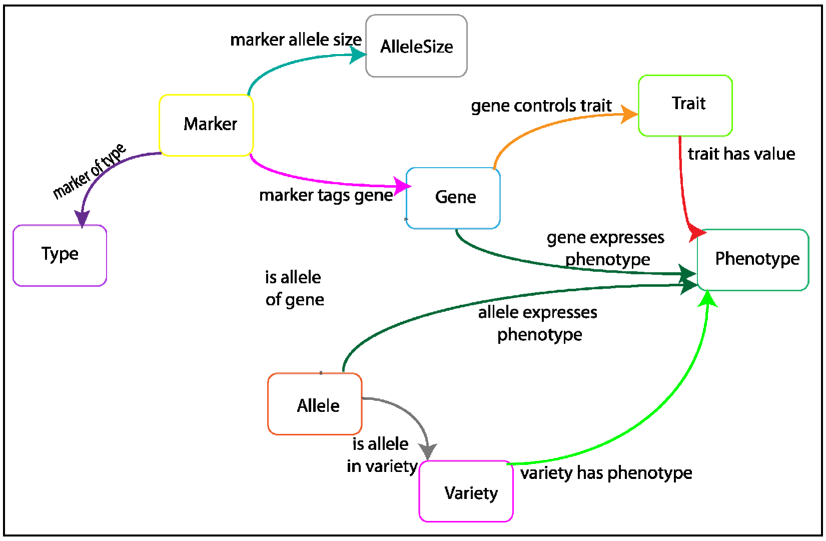


Fig. 1. a. Binary relations of the knowledge model for wheat MAS.

Binary relations may be used instead of ternary relations when an argument is missing. For instance, *marker_tags_gene* is used instead of *marker_tags_gene_in_variety* when the wheat variety is not mentioned in the text.

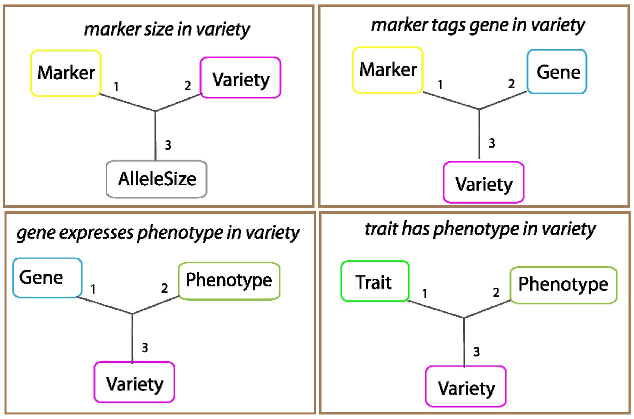
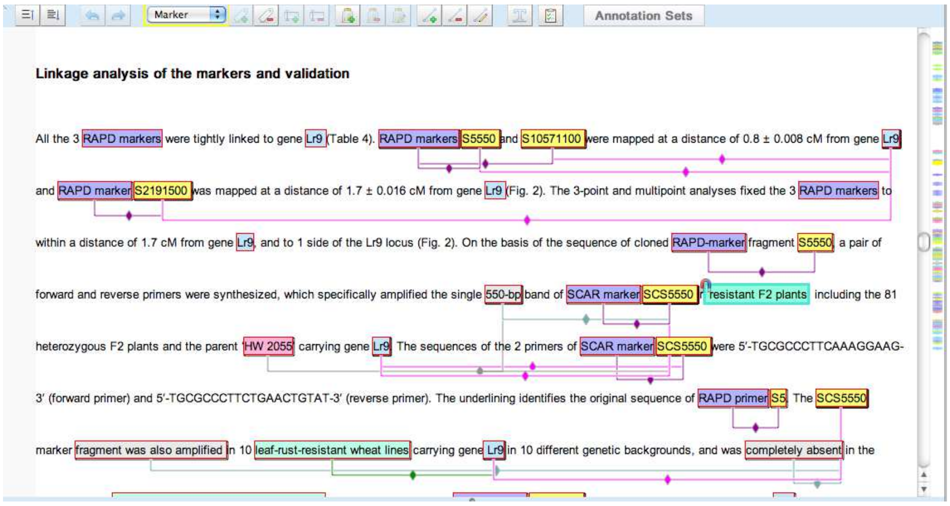


Fig. 1. b. Ternary relations of the knowledge model for wheat MAS.

Our RE method relies on the deep linguistic analysis and the supervised machine learning tools of the AlvisNLP pipeline [7]. The supervised machine learning method requires a training corpus of reference examples. For this purpose, we designed a corpus of 72 papers that were selected for their representativeness, most of them in *Theoretical and Applied Genetics, International Journal of Plant Breeding Research*. This journal publishes numerous manuscripts mentioning linkage between traits and molecular markers in wheat and is available on-line. 13 domain experts, mainly breeders and the GDEC authors of this paper annotated the corpus with the MAS knowledge model as annotation schema. To ensure effective and consistent annotation, we provided them with a guideline manual that describes the entities and the relations; it defines them and gives many examples that illustrate frequent and borderline cases. The annotation process follows the standard practice: double-blind annotation followed by adjudication. The text was automatically pre annotated by AlvisNLP with the entities that were frequent and easy to recognize in order to speed-up the manual annotation. The annotation editor AlvisAE [8] supported the whole process. The annotation campaign was defined by the annotation schema, the document collection and the two-step workflow. The 13 users were assigned a batch of documents to annotate and revise. We chose AlvisAE as annotation editor for its campaign specification tool and for its graphical user interface that was designed for non-computer scientists (Figure 2). The annotators were able to use it after one-day tutorial session. With AlvisAE the users annotated overlapping and discontinuous annotations. They also annotated co-references, which avoid the annotations of the repeated information.



Linkage analysis of the markers and validation

All the 3 RAPD markers were tightly linked to gene Lr9 (Table 4). RAPD markers S5550 and S10571100 were mapped at a distance of 0.8 ± 0.008 cM from gene Lr9 and RAPD marker S2191500 was mapped at a distance of 1.7 ± 0.016 cM from gene Lr9 (Fig. 2). The 3-point and multipoint analyses fixed the 3 RAPD markers to within a distance of 1.7 cM from gene Lr9, and to 1 side of the Lr9 locus (Fig. 2). On the basis of the sequence of cloned RAPD-marker fragment S5550, a pair of forward and reverse primers were synthesized, which specifically amplified the single 550-bp band of SCAR marker SCS5550 r 'resistant F2 plants including the 81 heterozygous F2 plants and the parent 'HW 2055' carrying gene Lr9. The sequences of the 2 primers of SCAR marker SCS5550 were 5′-TGCGCCCTTCAAAGGAAG-3′ (forward primer) and 5′-TGCGCCCTTCTGAACTGTAT-3′ (reverse primer). The underlining identifies the original sequence of RAPD primer S5. The SCS5550 marker fragment was also amplified in 10 leaf-rust-resistant wheat lines carrying gene Lr9 in 10 different genetic backgrounds, and was completely absent in the

Fig. 2. The main screen of AlvisAE annotation editor. The markers are highlighted in yellow, their types in purple, the varieties in pink, the genes in light blue and the allele size in grey. The lines figure the relationships between the entities. The type of the relationship determines the color of the line. For instance, the *maker tags the gene* relationship is in pink.

The double-blind annotation phase is achieved and the adjudication phase is on-going. The annotators fully annotated the 293 sections in the 72 corpus papers that were relevant. Table 1 displays the distribution of the entity and relation annotations. The distribution reflects the importance of the different information for breeding. The gene, variety, trait, marker and marker type are the most frequent and critical. Conversely, the alleles are rarely named, which explains the low number of allele annotations.

Table 1. Number of manual annotations in the wheat MAS training corpus.

Entities		**Binary relations**		**Ternary relations**	
Gene	1,826	marker_of_type	307	gene_expresses_phenotype_in_variety	103
Variety	1,284	gene_controls_trait	260	marker_size_in_variety	58
Marker	703	variety_has_phenotype	224	marker_tags_gene_in_variety	24
Type	508	marker_tags_gene	184	trait has phenotype in variety	24
Trait	603	allele_expresses_phenotype	107		**207**
Phenotype	403	is_allele_of_gene	107		
Alelle	368	is_allele_in_variety	64		
AlleleSize	153	marker_alleleSize	55		
	5 848	trait has value	34		
			1 342		

It does not affect the extraction results since the allele name is not required for the extraction of the linkage between the marker and the phenotype.

4 Named entity recognition

We used two different methods for the recognition of the named entities. We distinguished the rigid designators [9] from the other names. They are proper names, numbers and acronyms. They denote genes, markers, marker identification methods, allele sizes and varieties. Conversely, the phenotype and trait names are subject to more variation.

4.1 Recognition of proper names and acronyms

The NER method uses dictionaries such as gene and marker lists of the GrainGenes[1] database and hand-coded extraction patterns. The patterns identify typographic variations and perform word-sense disambiguation with the context of the target word. Disambiguation is particularly needed for the recognition of variety names that have frequent homonyms in the text, *e.g.* Leeds is cited both as a variety and a university. The quality of the method predictions has been evaluated with respect to the reference corpus. Table 2 displays the recall, precision and F_1 measures for an exact match and for a partial overlap between the predicted and the reference entities. F_1 is the harmonic mean of the precision and recall. The recognition of the names of the genes and markers is affected by homonymy: the marker of the gene is denoted by the same name as the gene. The reference annotations of markers and genes are mostly of good quality. Gain in prediction quality is to be found in the improvement of the disambiguation method and the gene name boundary identification. The performance of partial overlap in gene name recognition is 12 points over the exact match, which shows that the predicted gene name boundaries are often not correct.

Table 2. Quality of the named entity recognition.

	Exact match			**Partial overlap**		
	Recall	Precision	F_1	Recall	Precision	F_1
Gene	0,61	0,49	0,54	0,73	0,61	0,66
Marker	0,58	0,65	0,61	0,59	0,66	0,62
Type	0,54	0,62	0,58	0,56	0,64	0,60
AlleleSize	0,39	0,49	0,43	0,46	0,50	0,48

[1] http://wheat.pw.usda.gov/GG2/

The poor result of the recognition of the allele size is due to many errors in the reference corpus. Allele size names are numbers followed by *bp* (base pair) as *103 bp*. A close examination of the reference corpus revealed a high number of incorrect reference annotations, for instance *Ppd-D1a* designates an allele and not a size. Many annotated allele size represent an absence of the allele, for instance, *absence of PCR products*. An accurate correction of the reference corpus will allow a more significant evaluation of AlleleSize prediction quality.

The examination of the experimental results of named entity recognition shows us clear directions for further improvement.

4.2 Recognition of phenotypes and traits

3.2.1 The *ToMap* method

The recognition of the trait and phenotype terms cannot be efficiently achieved by the direct matching of dictionary entries with the text because of the high variability of the terms. Instead we used our ToMap method, previously named OntoMap [10]. ToMap requires a domain terminology and the results of a term extractor applied to the text. It matches the extracted terms with the terms of the terminology to determine which type of entity the extracted terms designate. The principle of the ToMap method is close to MetaMap method for the recognition of UMLS thesaurus terms in a corpus [11]. The matching relies on the similarity of the syntactic structures of the terms to be matched together. ToMap is applicable to any kind of text and terminology, being structured or not. ToMap has shown good results in the recognition of bacteria biotopes of the shared task *BioNLP-ST Bacteria Biotope* in 2011 and 2013 [12,13]. We used the term extractor BioYateA that is particularly well suited for this task because it provides the syntactic structure of the extracted terms and it extracts terms with prepositional phrases, e.g. *to crown rot* in *partial seedling resistance* <u>*to*</u> *crown rot* as described in [14].

3.2.2 The WheatPhenotype ontology

We built an ontology dedicated to the description of phenotypes and traits in wheat called *WheatPhenotype*. The available terminologies and ontologies were not fully relevant to our IE purpose. The most suitable is the Trait Ontology[2] (TO), a controlled vocabulary that describes traits of plants of agronomic interest. It includes relevant traits such as disease resistance, development traits, but also many traits that are irrelevant to wheat selection (*e.g.* biochemical, molecular, anatomy and morphology) and some over general trait (*e.g.* plant aspect) that are not mentioned in wheat selection texts.

[2] http://archive.gramene.org/db/ontology/search?id=TO:0000387

Conversely, TO lacks many specific traits and synonyms in all categories. The resistance to fungal disease is a critical trait in wheat selection. Only 8 fungus resistance concepts relevant to the wheat are defined in TO; we identified 24 in the texts and databases. Moreover, in the scientific papers disease resistance is often described by the resistance to the pathogens that cause the disease, for instance, *resistance to fusarium head blight* is equivalent to *resistance to Fusarium graminearum* where *Fusarium graminearum* is one of the fungus species that causes *fusarium head blight*. TO does not record such information. Moreover many different names can be used for each fungus. For instance, *Microdochium nivale* that also causes *fusarium head blight* is also called *Fusarium nivale*. The listing of all pathogen names and acronyms for all wheat disease is needed for efficient information extraction. Another important information is bread-making quality since the selected varieties determines the quality of the flour for bread making (mechanical and sensorial properties).

The current version of the *WheatPhenotype* ontology defines 409 concepts with 361 synonyms. Its hierarchical structure comprises 9 levels. Figure 3 shows the main levels.

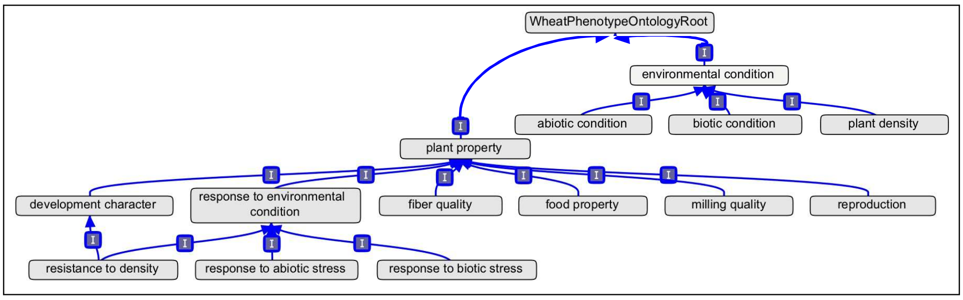


Fig. 3. The highest levels of the *WheatPhenotype* ontology.

The abiotic factors represent the physico-chemical conditions of the plant development (water availability, temperature, wind force, soil composition). The properties of the plant are organized in six sub trees, the response to environmental factors, the development, the reproduction, the product processing and the quality of the product (fiber and food). All traits and phenotypes are considered, including the less studied (*e.g.* aluminum tolerance), but the response to biotic factors, in particular to fungal and bacterial pathogens, is the most developed. The *cause* relationship links *pathogens* to the corresponding infectious *diseases* defined in the biotic condition sub tree. The *WheatPhenotype* ontology will be made publicly available in Obo format after alignment to TO. The common concepts with TO will be explicitly identified by a cross reference as an xref property.

3.2.3 Experimental results of phenotype and trait prediction

We applied the ToMap method to the terms extracted by BioYateA from the training corpus by using the WheatPhenotype ontology. As already noticed for the allele size annotation, the quality of the trait and phenotype annotation of the reference corpus was not sufficient for a reliable evaluation of ToMap predictions. The on-going adjudication phase will detect these errors and correct them.

(1) A frequent error confuses genotypes with phenotypes and alleles. For instance, the term *wild type* denotes an organism, but it is frequently annotated as a phenotype or as an allele, *e.g.* *wild type alleles WB357* means the allele WB357 of the wild-type line.

(2) The confusion between the environmental factors and the phenotypes is also frequent, for instance *winter* is annotated as a phenotype by analogy with the phenotype *winter habit* that denotes the growth period of the variety.

(3) The phenotype value is also confused with the trait, for instance *ToxA sensitivity* is confused with *ToxA sensitive*.

To obtain a reliable experimental result, we manually validated the results of the method applied on the abstracts of a subset of 870 papers. The ToMap method classified 299 terms as denoting a phenotype or a trait among the terms extracted by BioYateA. The manual validation of the classified terms yielded a precision rate but not a recall rate for which a reference annotation is required. Table 3 details the experimental results for two versions of the method, without and with disambiguation.

Table 3. Precision of the phenotype and trait prediction.

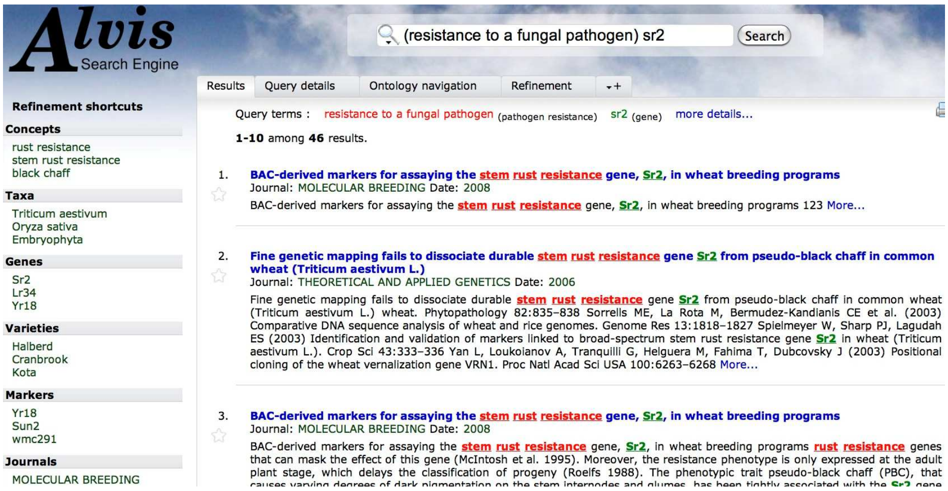
	Without disambiguation		**With disambiguation**	
Category	**# terms**	**Rate**	**# terms**	**Rate**
Positive	245	81%	212	95%
Correct and informative	227	76%	176	79%
Correct and general	18	6%	36	16%
Negative	54	19%	11	5%
Linguistic analysis error	5	1,7%	4	2%
ToMap error	16	5,4%	7	3%
ToMap setting error	33	11 %	0	0%

The first line gives a high precision rate of 81% that we divided into a correct and informative category (76%) and a correct but general category (6%). General terms are not useful for breeding but they are relevant for knowledge modeling. *Plant morphologic trait* is an example. A closer analysis of the false positive examples showed that a small number of the errors are due to linguistic preprocessing: word segmentation by the in-house tool SegMig and POS-tagging by TreeTagger [15]. Most of ToMap errors were due to an incorrect setting (11%). ToMap setting involves the setting of the list of term heads that are non-discriminant with respect to the named entity recognition goal. The list is dependent of the domain. The word *content* is an example of a non-discriminant head. It occurs in trait terms such as *Grain Protein Content* or *reduction in DON content*, but also in other terms, e.g. *polymorphism information content*. A better discrimination of terms with such heads was obtained by post-processing disambiguation hand-coded rules that used the words of the terms and their contexts. We also designed some domain specific rules to improve the boundary prediction by excluding irrelevant words such as *main* in the term *main growth habits*. As a result, the precision rate increased by 14 points to reach 95%. The total number of positive terms decreased from 245 to 212. Most of the 33 removed terms were not classified as negative, but merged with other terms as a result of the boundary correction (*main growth habits* and *growth habits* were counted as one instead of two in the previous setting).

These preliminary experiments are very promising yielding a precision rate of 95% in the prediction of traits and phenotypes for wheat selection. The reference manual annotations once consolidated will allow measuring the recall and F_1.

5 The *AlvisIR WheatMarker* bibliographical search engine

The semantic search engine *AlvisIR WheatMarker* indexes the document collection of 2,097 scientific papers about the linkage between molecular markers and genes that will be used for information extraction. The search engine index includes all entities defined in the knowledge model. The trait and phenotype index was built with the WheatPhenotype ontology, which means that query terms may contain high-level concepts of the ontology that will be searched together with all specializations and synonyms.



Refinement shortcuts

Concepts
rust resistance
stem rust resistance
black chaff

Taxa
Triticum aestivum
Oryza sativa
Embryophyta

Genes
Sr2
Lr34
Yr18

Varieties
Halberd
Cranbrook
Kota

Markers
Yr18
Sun2
wmc291

Journals
MOLECULAR BREEDING

Query terms : resistance to a fungal pathogen (pathogen resistance) sr2 (gene) more details...

1-10 among **46** results.

1. **BAC-derived markers for assaying the stem rust resistance gene, Sr2, in wheat breeding programs**
Journal: MOLECULAR BREEDING Date: 2008
BAC-derived markers for assaying the **stem rust resistance** gene, **Sr2**, in wheat breeding programs 123 More...

2. **Fine genetic mapping fails to dissociate durable stem rust resistance gene Sr2 from pseudo-black chaff in common wheat (Triticum aestivum L.)**
Journal: THEORETICAL AND APPLIED GENETICS Date: 2006
Fine genetic mapping fails to dissociate durable **stem rust resistance** gene **Sr2** from pseudo-black chaff in common wheat (Triticum aestivum L.) wheat. Phytopathology 82:835–838 Sorrells ME, La Rota M, Bermudez-Kandianis CE et al. (2003) Comparative DNA sequence analysis of wheat and rice genomes. Genome Res 13:1818–1827 Spielmeyer W, Sharp PJ, Lagudah ES (2003) Identification and validation of markers linked to broad-spectrum stem rust resistance gene **Sr2** in wheat (Triticum aestivum L.). Crop Sci 43:333–336 Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene VRN1. Proc Natl Acad Sci USA 100:6263–6268 More...

3. **BAC-derived markers for assaying the stem rust resistance gene, Sr2, in wheat breeding programs**
Journal: MOLECULAR BREEDING Date: 2008
BAC-derived markers for assaying the **stem rust resistance** gene, **Sr2**, in wheat breeding programs **rust resistance** genes that can mask the effect of this gene (McIntosh et al. 1995). Moreover, the resistance phenotype is only expressed at the adult plant stage, which delays the classification of progeny (Roelfs 1988). The phenotypic trait pseudo-black chaff (PBC), that causes varying degrees of dark pigmentation on the stem internodes and glumes, has been tightly associated with the **Sr2** gene

Fig 4. Interface of the semantic search engine *AlvisIR WheatMarker*

Figure 4 shows the results of the query (*resistance to a fungal pathogen*) *sr2* that aims at retrieving papers about *sr2* involvement in any resistance to fungal pathogens. The snippets (short extracts) of the 46 relevant documents are displayed below the query. The relevant terms are highlighted in the same colors as the query terms, *sr2* gene in green, resistance to fungal pathogen in red. With its semantic capabilities, AlvisIR retrieves many different fungal pathogen resistances such as *stem rust resistance* as highlighted in the three first snippets. The left panel displays the facets, the most frequent index values in the answer set. The query can be refined by the selection of a facet. *AlvisIR WheatMarker* semantic search engine is publically available[3]. The current version of the search engine does not index the relations for which the information extraction methods are under development.

Once the marker information will be fully extracted, it will be indexed by the *AlvisIR WheatMarker* search engine. It will also be integrated in a public database interconnected with all relevant genetic information, physical map, the 4000 known markers and the available wheat chromosome sequences [16,17,18]. It is worth to note that ToMap not only extract phenotypes and traits from the papers but also normalize them with respect to the WheatPhenotype Ontology, enabling heterogeneous data integration.

[3] http://bibliome.jouy.inra.fr/test/alvisir/FSOV/

6 Conclusion

The extraction of the available information on molecular marker published in scientific papers is a key issue for marker-assisted selection. It is particularly critical for wheat breeders that do not have access to this information in structured databases. We proposed a knowledge model that formalizes the knowledge needs in the form of an entity-relation schema. Our annotation framework involving a team of 13 breeders produces reference examples for training supervised machine learning methods and for the evaluation of prediction results. We proposed two methods based on linguistic analysis for the recognition of entities denoted by proper names and terms. The results evaluated on reference data yielded very encouraging results. The lack of structured vocabulary for extracting and normalizing phenotypes and traits led us to build the *WheatPhenotype* ontology. The prediction results and the *WheatPhenotype* ontology are used by the semantic search engine *AlvisIRWheatMarker* that index the full-text of the major papers of the domain. In the future, once consolidated, the reference wheat marker corpus will be made available to the community. It will be used for the training of the relation extraction methods. The overall approach will be then applied to other plants of agronomic interest, such as maize.

Acknowledgments

This work was partially funded by Oséo through the Quaero project and by FSOV through the SAM blé project. The authors thanks Jérôme Auzanneau (Agri-Obtention), Stéphane Boury (Caussade Semences), Emmanuelle Cariou-Pham (Arvalis), Clément Debiton (Unisigma), Noëmie Desmouceaux (Syngenta), Laure Duchalais (RAGT), Ellen Goudemand (Florimond Desprez), Pierre-Marie Le Roux (Secobra), Vanessa S. Windhausen (Saaten Union), Stephen Sunderwirth (Momont) for their participation to the annotation of the corpus.

References

1. Nédellec C: Learning Language in Logic – Genic Interaction Extraction Challenge. *Proc 4th Learning Language in Logic Workshop (LLL05)* 2005, pages 31-7, 2005
2. Hirschman L, Yeh A, Blaschke C, Valencia A: Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1, 2005.
3. Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J: Extracting bio-molecular events from literature – The BioNLP'09 Shared Task. *Computational Intelligence*, 27(4): 513-40, 2011.
4. Golik W, Dameron O, Bugeon J, Fatet A, Hue I, Hurtaud C, Reichstadt M, Salaün M.-C, Vernet J, Joret L, Papazian F, Nédellec C and Le Bail P-Y : ATOL: the multi-species livestock trait ontology. in proceedings of *The 6th Metadata and Semantics Research Conference (MTSR*

2012), pages 289-300. Springer Verlag Communications in Computer and Information Science Serie. Cadiz, Spain, 2012.

5. Collier N, Tran M-v, Le H-q, Ha Q-T, Oellrich A, et al.: Learning to Recognize Phenotype Candidates in the Auto-Immune Literature Using SVM Re-Ranking. *PLoS ONE* 8(10): e72965. doi: 10.1371/journal.pone.0072965, 2013.
6. Paux E, Faure S, Choulet F, Roger D, Gauthier V, Martinant J-P, Sourdille P, Balfourier F, Lepaslier M-C, Brunel D, Cakir M, Gandon B, Feuillet C: Insertion site based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. *Plant Biotechnol J*, 2009.
7. Nédellec C, Nazarenko A et Bossy R: Information Extraction. *Ontology Handbook*., S. Staab, R. Studer (eds.), Springer Verlag, Berlin, 2nd edition, pp 663-686, 2009.
8. Papazian F, Bossy R, Nédellec C. AlvisAE: a collaborative Web text annotation editor for knowledge acquisition. *Proc 6th Linguistic Annotation Workshop (The LAW VI),* pp 149-52, 2012.
9. Kripke, S: *Naming and Necessity*. Boston: Harvard University Press, 1982.
10. Golik W, Warnier P, and Nédellec C: Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. *Proc 9th Intl Conf Terminology and Artificial Intelligence (TIA 2011)*, pages 37-9, 2011.
11. Aronson A R, Lang F M: An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229-36, 2010.
12. Ratkovic Z, Golik W, Warnier P: Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. *BMC Bioinformatics*, 13(Suppl 11):S8, 2012.
13. Bossy R, Golik W, Ratkovic Z, Bessières P, and Nédellec C: BioNLP Shared Task 2013 – an overview of the bacteria biotope task. *Proc BioNLP Shared Task 2013 Workshop* 2013, pp 74-82. Association for Computational Linguistics (ACL), 2013.
14. Golik W, Bossy R, Ratkovic Z, Nédellec C: Improving term extraction with linguistic analysis in the biomedical domain. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'13),* Samos, Greece, 2013.
15. Schmid: Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.
16. Raats D, Frenkel Z, Krugman T, Dodek I, Sela H, Simková H, Magni F, Cattonaro F, Vautrin S, Bergès H, Wicker T, Keller B, Leroy P, Philippe R, Paux E, Doležel J, Feuillet C, Korol A, Fahima T: The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution. *Genome Biol*. Dec 20;14(12):R138. 2013.
17. Choulet F, Alberti A, Theil S, Glover N, Barbe V et al.: Analysis of the wheat chromosome 3B reference sequence reveals structural and functional compartmentalization. Science 345 DOI: 10.1126/science.1249721; 2014.
18. International Wheat Genome Sequencing Consortium (IWGSC) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345: DOI: 10.1126/science.1251788; 2014.