WikiParable

Data Categorisation Platform
Version 1.0

Cristina España-Bonet

November 16, 2015

Abstract

This document describes WikiParable, an on-line platform designed for data categorisation. Its purpose is twofold and the tool can be used both to annotate data and to evaluate automatic categorisations. As a main use case and aim of the implementation, the interface has been used within the TACARDI project to annotate Wikipedia articles in different domains and languages.[1]

[1]This work has been funded by the TACARDI project (TIN2012-38523-C02) of the Spanish Ministerio de Economía y Competitividad (MEC).

1 Motivation

Data categorisation or classification is the process of mapping data into categories. Humans tend to classify all their surroundings to ease the acquisition of information; and this is because classifications help to organise data –knowledge– and, therefore, to locate it in an efficient way, both to a human and to a machine.

Given a predefined taxonomy, items can be automatically assigned to a category. Usually, supervised machine learning algorithms are used for this purpose, but even in this case, a set of previously tagged items is necessary. When categories are well defined and exclusive, the task for a human with the adequate knowledge is easy but time-consuming, and not always a group of experts is at hand and can be used to do the annotation. This can be solved by resorting to crowdsourcing. Crowdsourcing is typically cheaper and faster than using a community of experts but the quality of the annotations can be damaged because of the lack of expertise.

This platform has been originally designed with the aim of proving the previous assertion. We pose several binary classification problems that will be annotated through the platform by two different communities of experts: astrophysicists and computer scientists. In a following study, the results obtained from this annotation will be compared to crowdsourced annotations.

Accordingly to the groups of experts, the categories chosen for the study are astronomy, computer science and sports. The elements to be tagged with an appropriate category are Wikipedia articles gathered by two different automatic systems implemented within the WikiTailor framework. One of the systems benefits from the user-made taxonomy of Wikipedia, and the other one uses standard information retrieval techniques to select a subset of articles related to the desired category from the complete Wikipedia. The second goal of this platform is to manually evaluate these systems.The data sets we make available to the platform have been selected to allow to do so as a by-product of the annotation.

In order to achieve the two goals, we have implemented an annotation platform for data categorisation customised to our specific problem. So, although it can be used for any classification problem, some pages are specific to our case. That includes the introductory pages and the analysis of the results. Next, Section 2 describes these pages and all the functionalities that the GUI has. The platform has been implemented in PHP with HTML and Java script embedded. All the data and annotations are stored in a MySQL database as described in Section 3. After the description of the two main components, we summarise the experience with the platform in Section 4.

2 The Web-based User Interface

As said in the previous section, the system and the web-based user interface have been implemented in PHP with HTML and Java script embedded. Currently, it comprises eleven screens where users can learn about the task, register, annotate or evaluate data sets, import or export them and see a first analysis of the results. The system accepts three kinds of users: (*i*) *unregistered users* which are only allowed to know about the task (definition and statistics), (*ii*) *registered users* with permissions to participate in the task and see their contributions, and (*iii*) *administrators* with additional permissions to upload/download data, see all users' contributions and interact with the database. The authentication and authorisation of users is controlled through Apache using htaccess.

Introduction to the task. The *front page* presents the general instructions for using the interface, from creating a user to categorise an article (Figure 1a). This page links to the *guidelines page* where some annotation criteria are established. The two pages fully depend on the specific task.

Registration of users. The Users drop-down menu in the navigation bar shows the functionalities related to the registration of users. The *registration page* (Figure 1b) allows to gather the relevant information about the user and the characteristics of its account. After submitting the form, the user receives a password that can be modified together with the other personal information in the *modify account page*.

Categorisation task. A drop-down menu in the *main page* displays all the data sets available for the annotation task (Figure 2a). Once the data set is selected with the Start button, an article and its possible categories is displayed (Figure 2b). Four actions are allowed on an article view: select the correct category and submit the choice, discard the current article to get a new one, undo the previous annotation, or submit the selected category and finish the current session.

Articles are shown to three different users in order to obtain reliable annotations. A data set is first annotated once and, when all the documents have one assignment, the second and third annotation rounds begin. This constraint intends to assure a complete set of annotations, at least with one value, for cases where there are not enough participants.

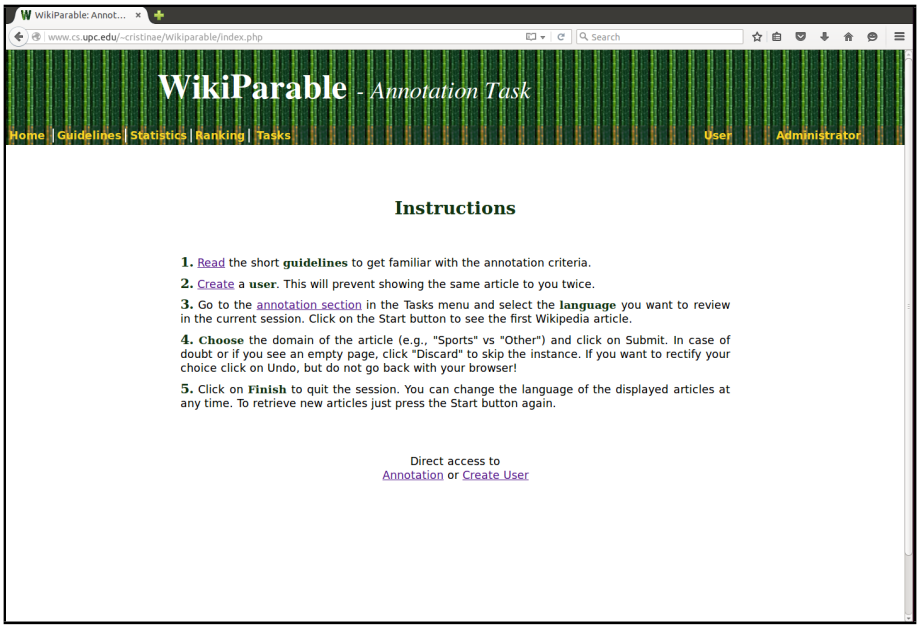
Statistics. Two pages report the figures on the task. The *statistics page* (Figure 3a) shows, for all the data sets active within the categorisation task, the number of annotations, the number of fully evaluated articles that that involves, the remaining number of annotations to finish the task for that data set, the number of users (annotators) that have contributed so far, and the Fleiss' kappa. The latter is an inter-annotator agreement score introduced in Ref. [1] that helps to interpret how reliable the annotations are (or equivalently, how difficult the task is).

The *rankings page* (Figure 3b) lists all the participants ranked according to the number of annotations they have done. The list is anonymous for the standard (non-administrator) users. A user can only see his/her name and an ID for the others. In our case, users belong to different groups and the first contributor from every group was promised a prize. The ranking marks the top annotator of each group with an identificative icon.

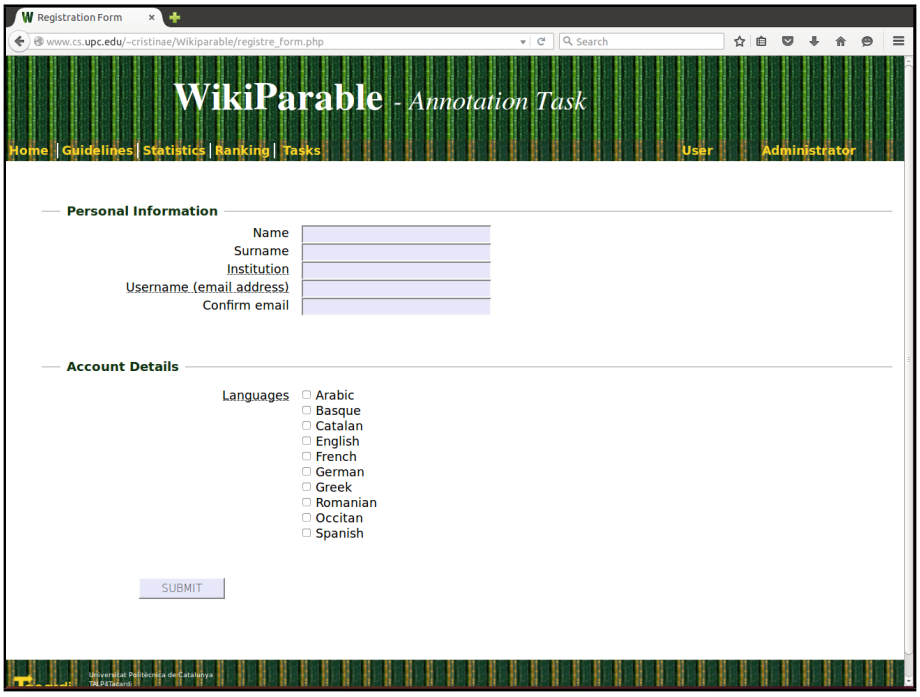
Administrator tasks. The Administrator drop-down menu in the navigation bar shows the functionalities related to a direct interaction with the database and a first analysis of the results in evaluation tasks.

The *import-export page* (Figure 4a) is an interface to upload or download the data sets into/from the database. The form is specific to the WikiTailor use case: the information describing the systems is collected through the form and a file with the list of Wikipedia articles' IDs can be uploaded. The same structure is used to download the results of the annotation for every data set. The file with the results is a cvs with the following format and content:

```
#IDarticle,IDdb,IDuser1,ann1,IDuser2,ann2,IDuser3,ann3,system,category,language
145600,291,5,1,4,1,2,1,1-100,astro,ca
756006,292,9,1,1,0,5,1,1-100,astro,ca
```

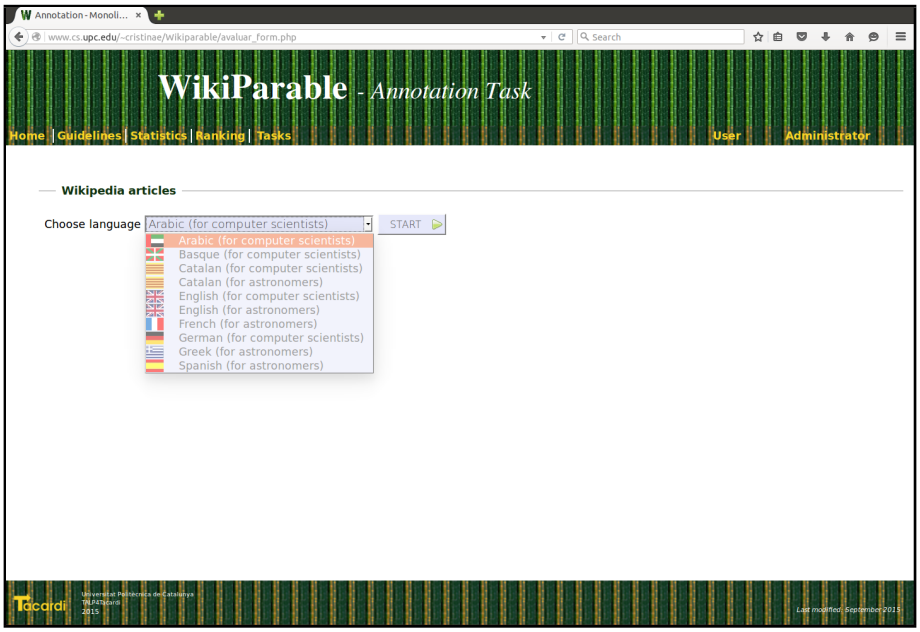


(a) Front page

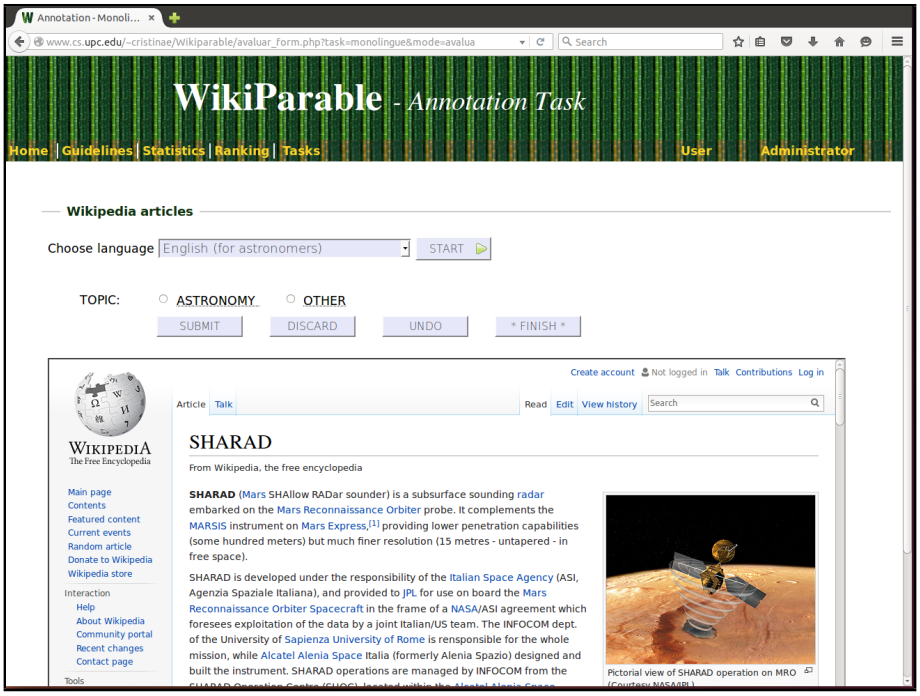


(b) Registration page

Figure 1: Screenshots of the introductory pages in the annotation interface

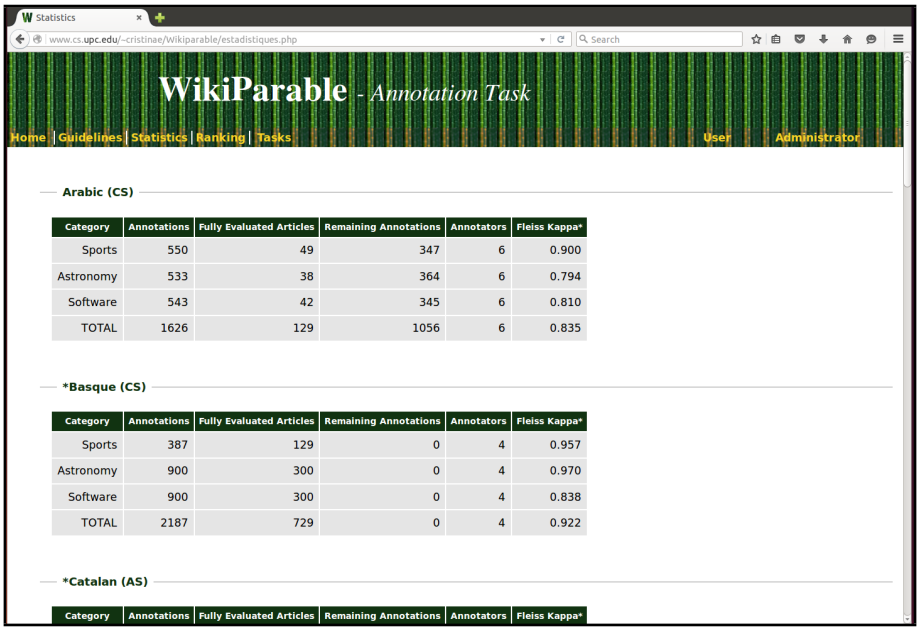


(a) Selection of the data set to categorise

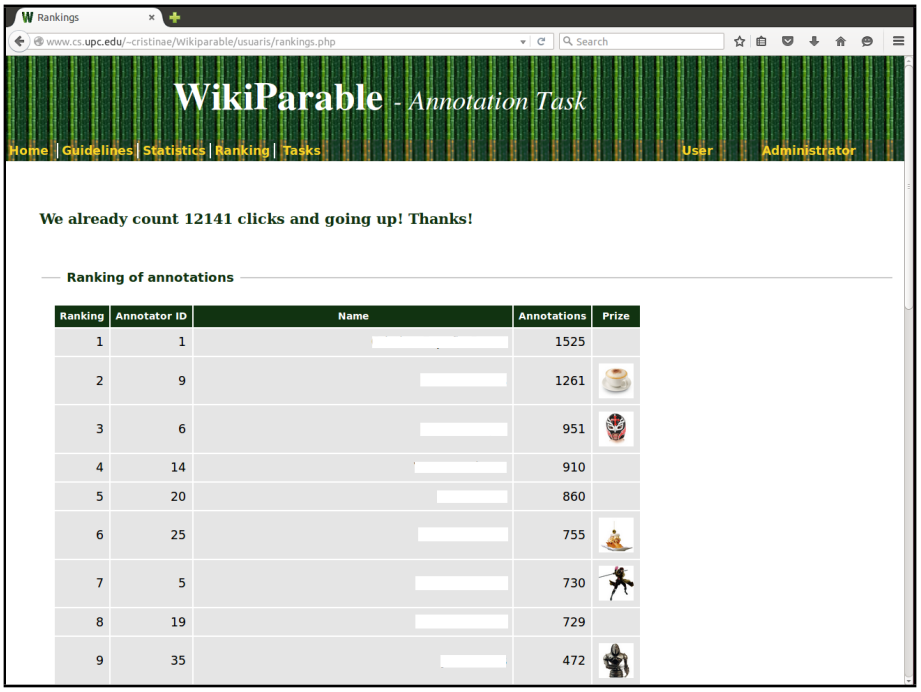


(b) Document view and annotation

Figure 2: Annotation main page in the interface

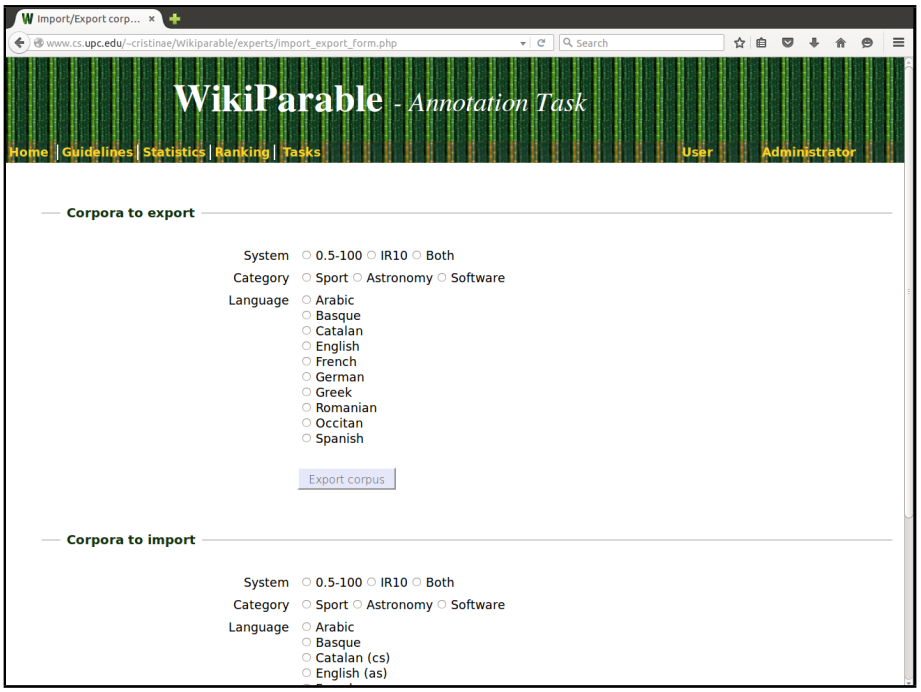


(a) Statistics for the available data sets

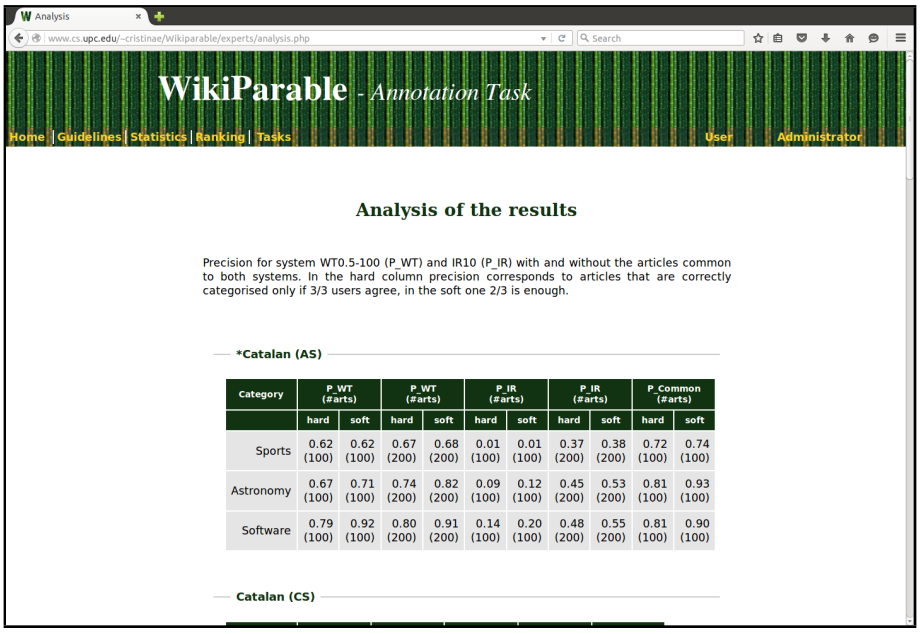


(b) Ranking of annotators (view for administrators where names have been removed)

Figure 3: Analysis of the ongoing annotation task



(a) Importing and exporting data sets into/from the database



(b) Precision for the different categorisation systems involved for the different data sets

Figure 4: Dealing with the results of the task

...

For cases where this platform is used to evaluate data sets, the *analysis page* reports the precision for every test set involved. Figure 4b shows, for several sets and subsets, the precision and the number of items it has been calculated on in two cases: (*i*) when there is full agreement among annotators (hard precision) and (*ii*) when an item is assigned to a category by the majority (2 out of 3) of annotators (soft precision).

3 The Relational Database

We use a MySQL database to store the data sets and annotations. The schema for our relational database including the tables, fields and foreign keys can be seen in Figure 5. In the following, we briefly describe the five tables:

Table 'user'. It contains the information about the user account: ID, name, surname, affiliation, kind of user and group, and tasks and languages to which he/she is subscribed.

Table 'article'. All the information related to an item is stored in table 'article', where an item can be any element to be categorised. For our use case, we include the id, domain and language of each Wikipedia article, a reference to the system that has selected it, a flag to identify articles that do no exist any more, and a boolean that indicates if that article has already all the necessary validations.

Table 'origen'. Describes the systems that have selected the items in 'article'. This data is only available for evaluation tasks.

Table 'verification'. This is the table that stores the results of each annotation. The fields of the table include the identifier for the item that has been validated, the user that has done it, the date of the validation, and the result of the validation. With our settings, every Wikipedia article must have three validations before the boolean field in 'article' blocks the item and prevents it from showing again.

Table 'statistics'. It contains the information about the interaction of the user with the interface. The system saves in this table the undos and discards that users do, with the information on the button that has been clicked, on which item, by which user and when.

3.1 Populating the Database

The database has been populated with 13 data sets of Wikipedia articles selected by two different systems, that implies 11670 rows in the table 'article'. The other tables, 'user', 'verification' and 'statistics', are populated dynamically during the course of the task.

4 Conclusions

We have developed an on-line platform for data categorisation. The platform is currently being used for data annotation and evaluation within the TACARDI project with two

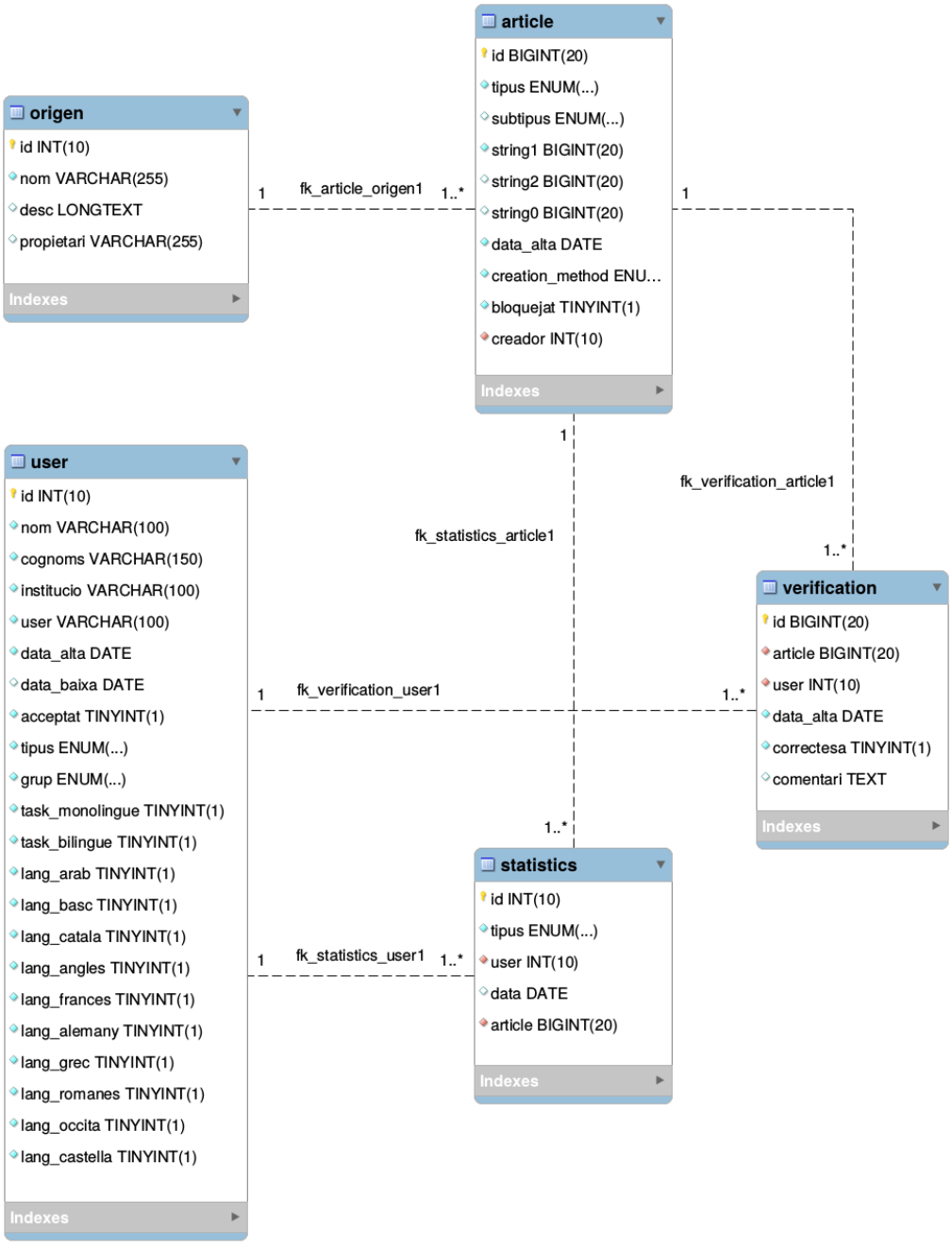


Figure 5: Schema of the MySQL database used in the application

purposes: (*i*) studying the relevance of the expertise of the participants in an *annotation* exercise and (*ii*) *evaluating* categorisations obtained by two automatic systems available within the WikiTailor framework.

The interface facilitates standard users to fulfil the annotation/evaluation task, while administrators can follow the status and quality of the categorisations and see a preliminary analysis of the results during the course of the experiment.

Although the platform has been specifically designed to chose among a set of possible categories for a Wikipedia article, WikiParable is easily extensible and can be adapted to any categorisation task.

References

[1] J.L. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.