Digital Humanities 2016
Kraków, 11–16 July

(http://dh2016.adho.org)

DH Home (http://www.dh2016.adho.org) / Abstracts (/abstracts/) / 19 (/abstracts/19)

Show info | How to cite | XML Version (/static/data/392.xml)

WordPress as a Framework for Automated Data Capture, Filtering and Structuring Processes. The New Order of the Authors

The Exhibitium Project

Note:

Generation of knowledge about temporary art exhibitions for a multivalent reuse was the topic of the proposal presented to the 2014 competition organized by the BBVA Foundation for projects in the field of Digital Humanities, resulting selected from over 250 submissions. The project website is available at: http://exhibitium.com. The Exhibitium Project began in January 2015 and will end in December 2016, so currently we are completing the first phase.

, awarded by the BBVA Foundation, is a data-driven project developed by an international consortium of research groups.

Note:

They are: iArtHis_Lab (http://www.iarthislab.es) and Khaos (http://khaos.uma.es) at the University of Málaga; Techne, ingeniería del conocimiento y del producto (http://www.ugr.es/~tep028/quienes_somos_es.php (http://www.ugr.es/~tep028/quienes_somos_es.php)) at the University of Granada; and CulturePlex at the University of Western Ontario (http://www.cultureplex.ca).

One of its main objectives is to build a prototype that will serve as a base to produce a platform for the recording and exploitation of data about art-exhibitions available on the Internet.

Note:

Specifically, the ultimate Exhibition's purpose is to extract unprecedented and strategic knowledge about temporary art exhibitions through the use of a variety of data mining techniques.

Therefore, our proposal aims to expose the methods, procedures and decision-making processes that have governed the technological implementation of this prototype, especially with regard to the reuse of WordPress (WP) as a development framework.

According to the project's purpose, it was necessary to create a device that, to the extent possible, could capture in automated way information on art exhibitions from any Internet source. Consequently, the inquiry into the possibilities of web mining strategies emerged as a priority from the early stages. Taking into account the high expressiveness and flexibility of linguistic structures usually used in the description of art exhibitions, our project opted for a mixed platform which combines the potential modeling system based on textual indicators, the heuristic means that characterize some methods -such as the Bayesian classification- and the human supervision provided by a well trained team of editors.

1. 1. General overview. WordPress as a framework of the Expofinder system

As Baumgartner et al. (2009: 1) established, the web data extraction task is usually divided into five different functions: (1) the web interaction, which mainly comprises the navigation through predetermined web pages containing the information sought; (2) the extraction of the searched data by means of a software that identifies, extracts and transforms them into a structured format; (3) the setting of a specific calendar that enable to perform automatically the extraction tasks in regular sequence; (4) the processing of the captured data, which includes filtering, transformation -if applicable-, refining and integration; and (5) the delivery of the resulting structured data to a variety of data analysis-based systems.

Assuming this distribution as the most convenient for our purposes, we decided to include them in the Exhibitium's architecture grouped into two large blocks.

A. A block consisting of an automated capture system of information robust enough to ensure the reliability of the collected data.

B. A second block made up by the set of elements necessary to store the data, including functions for filtering, cleaning, management, structuring and description. This block also incorporates a system to export the collected data to those platforms that will process and analyze them during the second phase of the project.

Block A was called Beagle, and block B became known as ExpoFinder. Both blocks work in a coordinated manner, so that what is extracted by Beagle is put at the disposal of ExpoFinder. The two blocks are part of an unified system configured by a cyclic algorithm: Beagle captures, ExpoFinder analyzes and approves the captured information, the team of editors validates or discards what ExpoFinder has previously approved, and Beagle recaptures again (see figure 1).

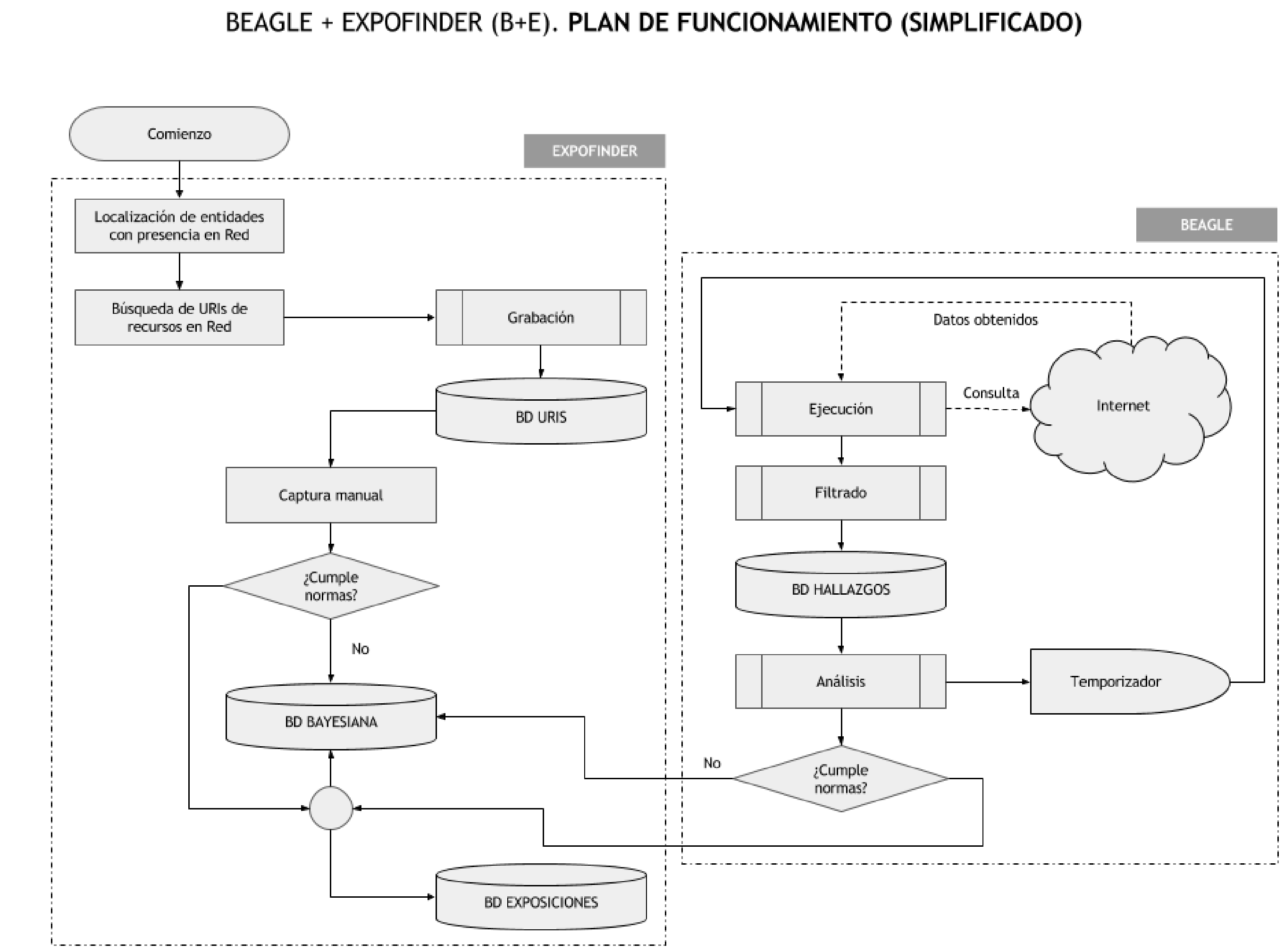


Figure1. Beagle + ExpoFinder. Operating plan (simplified)

Regarding the software, after preliminary versions based on own developments, it was decided that the most interesting option between the free software and open source solutions currently available (as the openness philosophy is a sine qua non requirement of this project) would be to use WP as framework of the system.

Note:

> Although, in reality, according to the Tom McFarlin's statement in his popular page «tuts +» (http://tutsplus.com/), it is more a foundation that a framework. And maybe he is right: a framework consists of a set of conventions as well as libraries and tools that allow us to easily start working on an application. In short, it provides the means by which an application can be built from scratch, from the database schema to the front end. However, a foundation allows to «extend» an existing application. WP has its own well-defined internal mechanisms, and the foundation simply expands its operation or takes advantage of it for their own benefit.

The main benefits that the use of WP as framework offers for our project can be synthesized in the following items: a database with a flexible and very solid organizational structure; a layer of a core application with numerous hooks which allow to maximize its functionality; and a high easy management system to carry out tasks on the two sides (server and client), assuming both administrator and user roles.

> *Note:*
>
> The advantages that a robust mechanism as such provided by WP offers for the maintenance of a security system (essential in any development accessible through Internet), or the substantial savings in time and resources involved in a CRUD structure records management -which is both sufficiently malleable to suit any need and rigid enough to follow canonical deployment patterns (such as the «nonce» safeguards in the capture forms), are weighty arguments when opting for the use of one framework or another.

Thus, for the implementation of the Beagle-ExpoFinder system we took advantages of the predefined data base, the available APIs and the set of data visualization templates to build solutions using an application that is already fully functional.

We used WP without adaptations, that is, as it can be downloaded from the Internet. All the functionality of our application lies, then, on the code itself that constitutes WP, so it is not supported on variant versions (forks) of the original program. Hence, any improvement provided by the computing community will be directly usable by our project without further adaptations. As part of the requirements of the development of the Beagle-ExpoFinder system (B + E), from the beginning it was considered that the programming work did not constitute a «tailored suit» for the Exhibitium project. On the contrary, we expect that this work can be useful in other projects with a small number of modifications or by using configuration files or other similar systems. For that reason, our choice was to implement B + E by means of a «WP theme», solution that easily allows us to readapt the software to different purposes.

2. 2. Beagle and Expofinder development and technical features

Beagle captures – as it has been said- by automated means web data concerning temporary art exhibitions from any source of information, and includes a filtering mechanism. The automated capture process uses the tools of WP API, particularly WPCron. Likewise, the frequency of the process is configured according with the options offered by ExpoFinder to the system administrator. Beagle employs two statistical complementary functions to «predict» the degree of the adequacy of the captured information to the ExpoFinder preconditions:

Note:

Even though this document is not largest enough to expose in detail the list of selected preconditions of significant terms used by Beagle in order to filter the captured information, we want to emphasize that this is a «weighted» relationship of lexemes in which each root term is assigned a total «weight» in the set (1 to 3). When the entire process is complete, the absolute amount of the sum corresponding to the found terms is weighted with the relative values (relating to the length of the text where they have been detected) to assign a positive or negative evaluation to the whole information.

1. One of this is based on the intersection of a set of «positive» and «negative» keywords with a proportional weight assigned to each one, which is also based on the shortest path algorithm of Bellman-Ford;

Note:

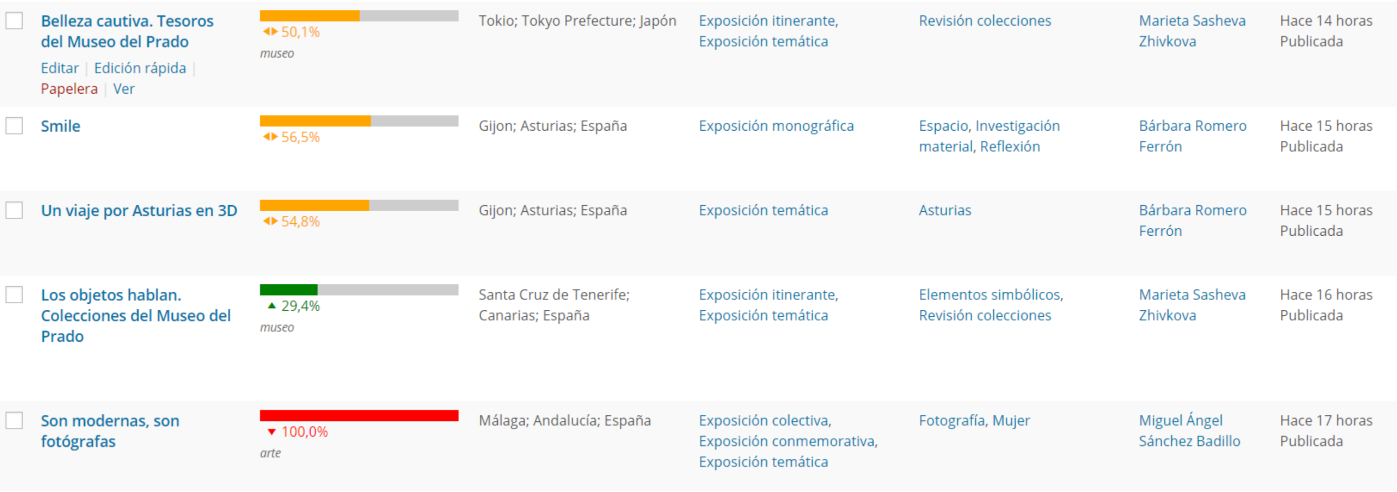
The Bellman-Ford algorithm (or Bellman-Ford-Moore) calculates the shortest paths from a single source vertex to all other vertices in a weighted digraph. It is slower than Dijkstra's algorithm for the same problem, but more versatile, since it is suitable to deal with graphs using negative numbers for edge weights. ExpoFinder takes advantage of it in its weighting mechanism, useful for us because we work with lists of lexemes for words used as «positive» or «negative» markers.

2. The other is defined by its heuristic nature; it employs a naive Bayesian classifier

Note:

In machine learning terminology, the «naive Bayesian» constitutes a family of simple probabilistic classifiers based on the application of Bayes' theorem about the hypothesis of independence between variables. Widely studied since the 1950s, it began to be used since the beginning of the next decade as a taxonomy method capable of self-optimization in the recovery community text. We use the frequency of occurrence of a given lexeme as a trigger, so that ExpoFinder can contribute to the semi-automated selection of relevant information from the experience gained. It is not a pure discriminative mechanism, but an auxiliary tool that has proven to be useful for application operators.

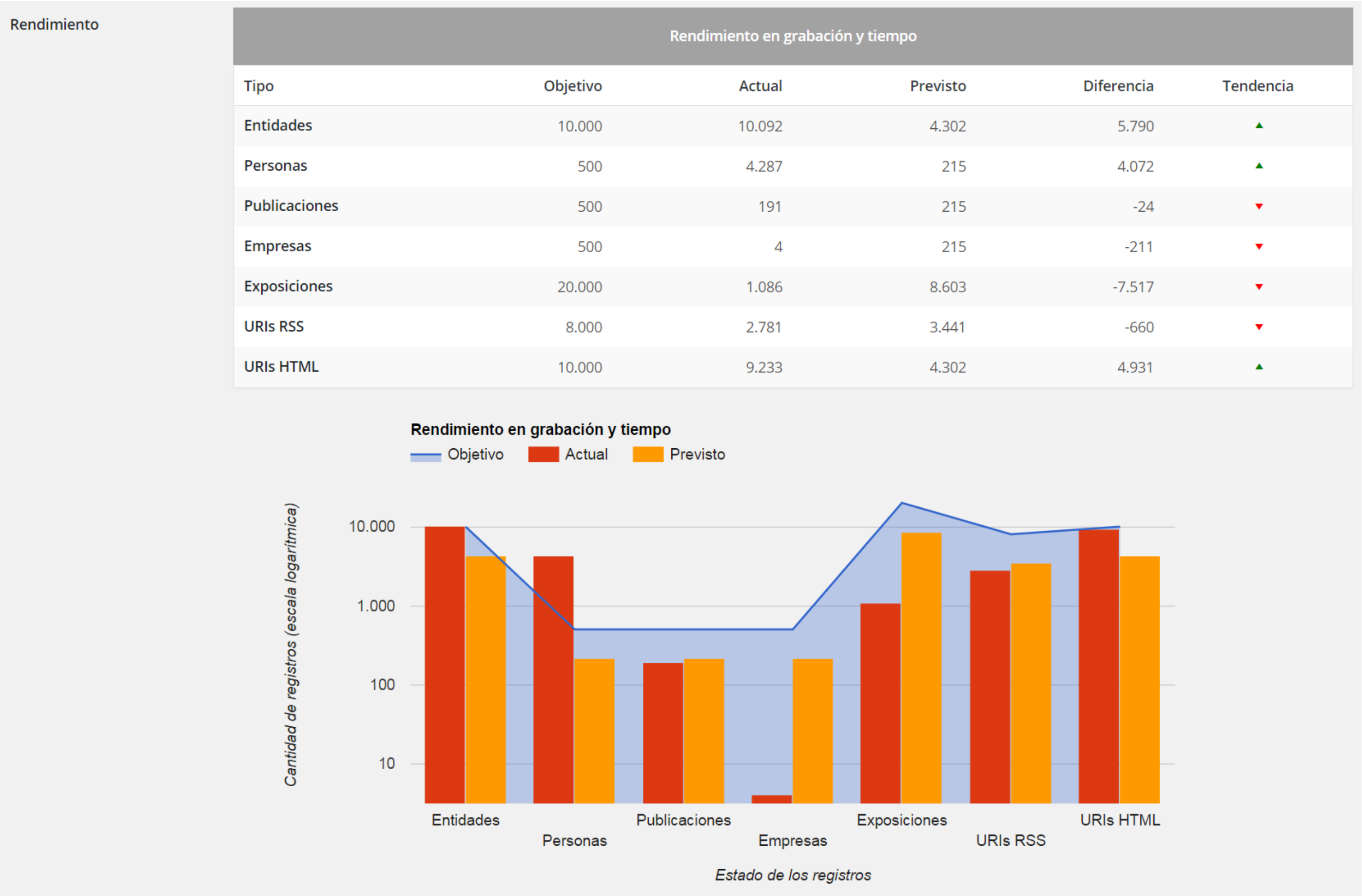
to guide the «human» editor during the task of discriminating whether or not an information captured by Beagle is valid. The latter is able to improve their efficiency through continuous learning processes (each discarding or acceptance made by the «human» editor refines the system «perceptiveness» (see figures 2 and 3).



□	Belleza cautiva. Tesoros del Museo del Prado Editar Edición rápida Papelera Ver	◂▸ 50,1% museo	Tokio; Tokyo Prefecture; Japón	Exposición itinerante, Exposición temática	Revisión colecciones	Marieta Sasheva Zhivkova	Hace 14 horas Publicada
□	Smile	◂▸ 56,5%	Gijon; Asturias; España	Exposición monográfica	Espacio, Investigación material, Reflexión	Bárbara Romero Ferrón	Hace 15 horas Publicada
□	Un viaje por Asturias en 3D	◂▸ 54,8%	Gijon; Asturias; España	Exposición temática	Asturias	Bárbara Romero Ferrón	Hace 15 horas Publicada
□	Los objetos hablan. Colecciones del Museo del Prado	▲ 29,4% museo	Santa Cruz de Tenerife; Canarias; España	Exposición itinerante, Exposición temática	Elementos simbólicos, Revisión colecciones	Marieta Sasheva Zhivkova	Hace 16 horas Publicada
□	Son modernas, son fotógrafas	▼ 100,0% arte	Málaga; Andalucía; España	Exposición colectiva, Exposición conmemorativa, Exposición temática	Fotografía, Mujer	Miguel Ángel Sánchez Badillo	Hace 17 horas Publicada

Figure 2. Screenshot. Exhibitions list (fragment). See the Bayesian classifier indicator

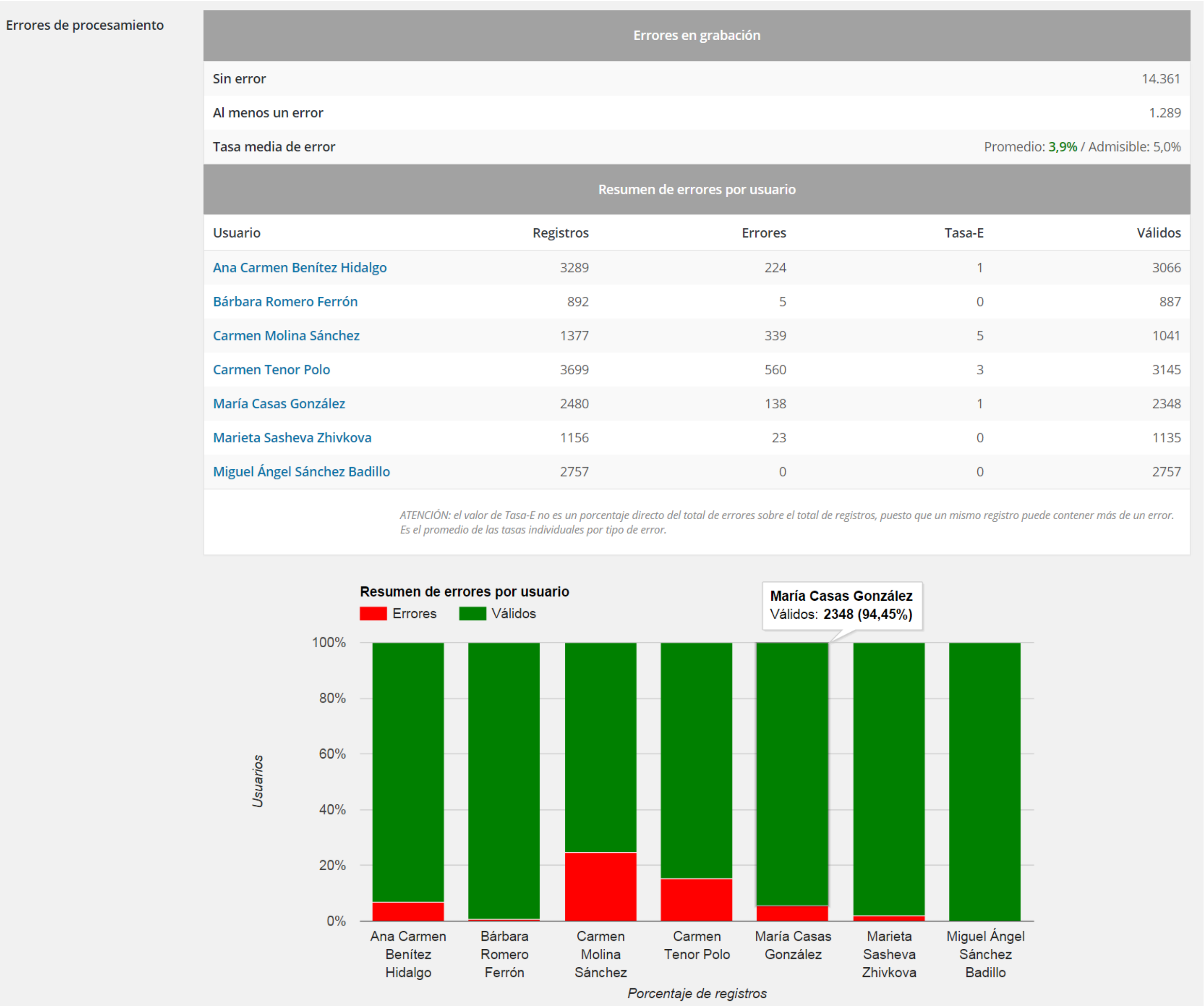
ExpoFinder also includes a control system (QC) that identifies the mistakes and failures, which are also associated with the human editor who made them, so that he/she can perform the appropriate corrections (see figures 3 and 4).



Rendimiento

Rendimiento en grabación y tiempo					
Tipo	Objetivo	Actual	Previsto	Diferencia	Tendencia
Entidades	10.000	10.092	4.302	5.790	▲
Personas	500	4.287	215	4.072	▲
Publicaciones	500	191	215	-24	▼
Empresas	500	4	215	-211	▼
Exposiciones	20.000	1.086	8.603	-7.517	▼
URIs RSS	8.000	2.781	3.441	-660	▼
URIs HTML	10.000	9.233	4.302	4.931	▲

Figure 3. Screenshot. Automated evaluation of efficiency



Errores de procesamiento

Errores en grabación	
Sin error	14.361
Al menos un error	1.289
Tasa media de error	Promedio: 3,9% / Admisible: 5,0%

Resumen de errores por usuario

Usuario	Registros	Errores	Tasa-E	Válidos
Ana Carmen Benítez Hidalgo	3289	224	1	3066
Bárbara Romero Ferrón	892	5	0	887
Carmen Molina Sánchez	1377	339	5	1041
Carmen Tenor Polo	3699	560	3	3145
María Casas González	2480	138	1	2348
Marieta Sasheva Zhivkova	1156	23	0	1135
Miguel Ángel Sánchez Badillo	2757	0	0	2757

ATENCIÓN: el valor de Tasa-E no es un porcentaje directo del total de errores sobre el total de registros, puesto que un mismo registro puede contener más de un error. Es el promedio de las tasas individuales por tipo de error.

Figure 4. Screenshot. Quality control (QC). Resume

In its current state of development, the Beagle-ExpoFinder system captures and selects daily about 100 references from more than 12,000 web sources of information. Its error rate during the recording and validation processes is about 3.9%, below the 5% initially considered as permissible.

Bibliography

1. **Baumgartner, R. et al.** (2009). Web Data Extraction System. In *Encyclopedia of Database Systems*. Springer-Verlag.
2. **Kokkoras, F. et al.** (2013). DEiXTo: A Web Data Extraction Suite. *Proceedings of 6th Balkan Conference in Informatics* (BCI 2013), Nueva York: ACM, pp. 9-12.
3. **Pree, W.** (1994). *Design Patterns for Object-Oriented Software Development*. Reading, Massachussets, USA: Addison-Wesley, ACM Press Books.
4. **Raposo, J. et al.** (2002). The Wargo System: Semi-Automatic Wrapper Generation in Presence of Complex Data Access Modes. *Database and Expert Systems Applications*. *Proceedings*, 13th International Workshop, IEEE, pp. 313-17.