Inf Syst E-Bus Manage (2013) 11:569–595
DOI 10.1007/s10257-012-0209-5

ORIGINAL ARTICLE

Using statistics, visualization and data mining for monitoring the quality of meta-data in web portals

Marcos Aurélio Domingues · Carlos Soares · Alípio Mário Jorge

Received: 2 April 2012 / Revised: 6 October 2012 / Accepted: 16 November 2012 /
Published online: 9 December 2012

Abstract The goal of many web portals is to select, organize and distribute content in order to satisfy its users/customers. This process is usually based on meta-data that represent and describe content. In this paper we describe a methodology and a system to monitor the quality of the meta-data used to describe content in web portals. The methodology is based on the analysis of the meta-data using statistics, visualization and data mining tools. The methodology enables the site's editor to detect and correct problems in the description of contents, thus improving the quality of the web portal and the satisfaction of its users. We also define a general architecture for a system to support the proposed methodology. We have implemented this system and tested it on a Portuguese portal for management executives. The results validate the methodology proposed.

Keywords Web data analysis · Meta-data quality · Quality of process · Content management · Web portals

This paper extends and details the work presented in Soares et al. (2005) and Domingues et al. (2006).

M. A. Domingues (✉)
INESC TEC, INESC Technology and Science (formerly INESC Porto), Porto, Portugal
e-mail: maddomingues@gmail.com; marcos.a.domingues@inescporto.pt

C. Soares
INESC TEC (formerly INESC Porto) and Faculty of Economics,
University of Porto, Porto, Portugal
e-mail: csoares@fep.up.pt

A. M. Jorge
LIAAD/INESC TEC and Faculty of Sciences, University of Porto,
Porto, Portugal
e-mail: amjorge@fc.up.pt

1 Introduction

The aim of many web portals is to select, organize and distribute content (information, or other services and products) in order to satisfy its users/customers. The methods to support this process are to a large extent based on meta-data (such as keywords, categories, authors and other descriptors) that describe content and its properties. For instance, search engines often take into account keywords that are associated with the content to compute their relevance to a query. Likewise, the accessibility of content by navigation depends on their position in the structure of the portal, which is usually defined by a specific meta-data descriptor (e.g., category).

Meta-data is usually filled in by the authors who publish content in the portal. The publishing process, which goes from the insertion of content to its actual publication on the portal is regulated by a workflow. The complexity of this workflow varies: the author may be authorized to publish content directly; alternatively content may have to be analyzed by one or more editors, who authorize its publication or not. Editors may also suggest changes to the content and to the meta-data that describe it, or make those changes themselves.

In the case where there are many different authors or the publishing process is less strict, the meta-data may describe content in a way which is not suitable for the purpose of the portal, thus decreasing the quality of the services provided. For instance, a user may fail to find relevant content using the search engine if the set of keywords assigned to it are inappropriate. Thus, it is essential to monitor the quality of meta-data describing content to ensure that the collection of content is made available in a structured, inter-related and easily accessible way to the users.

In this paper we describe a methodology to monitor and control the quality of the meta-data used to describe content in web portals. We assume that the portal is based on a Content Management System, which stores and organizes content using a database. The core of this methodology is a system of metrics, that objectively assess the quality of the meta-data and indirectly the quality of the service provided by the portal. The metrics are defined on data that is collected by the Content Management System (the meta-data) and by the portal's web servers (the access logs). We also define a general architecture for a system to support the proposed methodology. We expect that the deployment of this methodology will support the publishing process with an increased quality of meta-data. For instance, it may be possible to publish content as soon as it is inserted. The editor can make corrections to the meta-data later on, based on the reports provided by the system. This methodology was first introduced in an earlier paper (Soares et al. 2005), where preliminary ideas were described. In this paper we present a consolidated version of the methodology and the system that implements it, describing them in significantly greater depth.

As part of this work, we have extended the framework for assessing the quality of Internet services, which was defined by Moorsel (2001). We have also adapted the general set of data quality dimensions proposed by Pipino and Wang (2002) to the particular case of content meta-data. Technically, we have adapted the data warehouse proposed by Domingues et al. (Domingues et al. 2007)

to be the repository of information to our system. Finally, we have successfully applied the methodology to PortalExecutivo, a Portuguese portal for management executives.

This paper is organized as follows. We start by presenting related work (Sect. 2) In Sect. 3 we describe the methodology to monitor the quality of content meta-data. In Sect. 4 we detail the architecture of the system that implements this methodology. We then present a case study (Sect. 5) and conclude by summarizing the work and suggesting a few of the possible lines of future work (Sect. 6).

2 Related work

The evaluation of the quality of Internet services using quantitative metrics is discussed by Moorsel (2001). Three different quality levels are considered: Quality of Services (QoS), Quality of Experience (QoE) and Quality of Business (QoBiz). The author argues that the focus must be on QoE and QoBiz metrics. These attempt to quantify the quality of the service from the perspective of the user (e.g., response time of the site) and of the business (e.g., number of transactions), respectively, while QoS metrics are concerned with the platform (e.g., availability of the site). Moorsel also introduces a framework to evaluate the QoBiz, motivated by the emergence of three types of services that impact quantitative evaluation: business to consumer services, business to business services, and service utility through service providers.

Most of the work on the quality of Internet services focuses on one QoE dimension, usability. The usability of web sites has been recently regarded as an interesting ground for data analysis and data mining techniques. One of the first systems for monitoring the usage of web sites was proposed in Zaïane et al. (1998). The system, called WebLogMiner, combines OLAP (OnLine Analytical Processing) with data mining techniques (association, prediction, classification, time-series analysis, etc.) in a multidimensional data cube to predict, classify, and discover interesting patterns and trends. A more recent system was proposed in Velasquez and Palade (2008). The system applies clustering algorithms on web data stored in the data warehouse in order to recommend hyperlinks to be added to or eliminated from the current site, and (key)words to be used as "words to write" in the current and future pages. In Spiliopoulou and Pohle (2001), the authors propose a number of objective metrics to assess the usability of web sites. These metrics take into account users accesses and knowledge about the site. Das and Turkoglu help the web designer and web administrator to improve the impressiveness of a web site by using path analysis techniques to determine occurred link connections on the web site (Das and Turkoglu 2009). Visualization tools have also been used to identify difficulties in web site navigation (Berendt 2002; Cadez et al. 2003). A study that measures the quality of web content and not the usability of the sites is the work of Baeza and Rello (2012), who propose a measure for estimating the lexical quality of the web content, i.e., the representational aspect of the textual web content.

It is widely agreed that the quality of meta-data is important for the quality of Internet services and particularly in digital content-based services, such as digital

libraries (Park 2009; Ochoa and Duval 2009; Nichols et al. 2008). The ability to find the desired information heavily depends on the quality of the meta-data describing it (e.g., keywords or categories). However, the problem has not been extensively studied (Park 2009; Ochoa and Duval 2009). This is particularly surprising given that there is abundant evidence that meta-data collected manually or automatically is of low quality (Isinkaye et al. 2012; Lex et al. 2012; Blanco et al. 2011). One of the difficulties is in defining the concept of meta-data quality. Therefore, very general definitions, such as the degree to which the meta-data supports "the functional requirements of the system it is designed to support" or, in other words, "quality is about fitness for purpose" (Guy et al. 2004).[1] Despite its generality, this is the definition considered in this work, as it is in line with the concept of internet service quality presented earlier.

In this work, we assume that meta-data concerns data that characterizes other data. According to Isinkaye et al. (2012), there exist three types of meta-data: administrative, descriptive and structural. Administrative meta-data provides information that helps in tracking and managing information resources, as for example, copyright and licensing information. Meta-data which is classified as descriptive provides a way of identifying and describing an information resource to facilitate searching and retrieval (e.g., list of keywords). Structural meta-data is used to describe the organization of the content of an information resource in order to tie each resource to the other to make up logical units (e.g., content category). In this work, we will focus on descriptive and structural meta-data, although it is clear that good quality administrative meta-data can also be useful to improve the quality of Internet services. However, it should also be noted that integrating the latter kind of meta-data in the approach proposed here is trivial.

The traditional approach to assess the quality of meta-data has been the manual review of samples of meta-data. While this kind of analysis can be useful for small-sized and slow-growing sets of meta-data, it becomes impractical for large ones. To handle with this issue, quality metrics have been proposed and computed automatically in order to provide an estimation of quality for large sets of meta-data (Ochoa and Duval 2006). For instance, in Vuong et al. (2007) the authors study the problem of maintaining meta-data for open web content. As web meta-data maintenance involves manual efforts, they propose to reduce the efforts by introducing the Key element-Context (KeC) model to monitor only those changes made on web page content regions that concern meta-data attributes while ignoring other changes. They also develop evaluation metrics to measure the number of alerts and the amount of efforts in updating meta-data. Another example, in Nichols et al. (2008) the authors describe a web based meta-data quality tool that provides statistical descriptions and visualizations of Dublin Core meta-data harvested via the OAI protocol. As a final example, in Isinkaye et al. (2012) a set of rules and norms are proposed in order to guarantee the integrity and the consistent usage of meta-data. The rules and norms define in what form and with what logical constraints meta-data fields can be filled. The categorization of content into an adequate taxonomy is essential for the quality of the service provided by a web site.

[1] As mentioned in Park (2009).

Taxonomies can be used by powerful interfaces for disclosing large information repositories. In Fluit and Wester (2002), the authors present a number of ways in which the Cluster Map, a component for the visualization of instantiated taxonomies, can help a user gains insight in the information, detects anomalies, monitors the information as it evolves over time and assesses the quality of the output of automatic document classification tools. The proposed visualizations are presented in the context of one of their customers, for which they create web portals based on taxonomies, providing access to a large document collection. One important shortcoming of all these approaches except for the latter is the lack of flexibility of the tools proposed. They enable the user to identify meta-data quality problems but they provide limited functionalities to investigate the cause of those problems (e.g., which document or author is responsible for the problem; are there subjects which are more prone to problems than others?). Additionally, most of the literature focuses on Digital Libraries, Federated Collections and Learning Object Repositories (Bruce and Hillmann 2004; Stvilia et al. 2004; Ochoa and Duval 2006, 2009; Park 2009), which are somewhat different from the e-business domain addressed in this work.

The study of meta-data quality is highly related to the field of data and information quality (Park 2009). This field deals with general methodologies such as the one proposed by Pipino and Wang (2002), which is the basis for the current work. In this methodology, the authors propose a set of general principles for developing metrics to assess the quality of data. The metrics are organized according to 16 dimensions, ranging from *free-of-error*, representing metrics that account for incorrect values, to *value-added*, representing metrics that account for the benefits brought by the data. They also describe three functional forms which are often used to develop data quality metrics, namely simple ratios, min/max and weighted averages. They illustrate these principles with two case studies, including both subjective and objective metrics.

The work presented in this paper shares some characteristics with most of the approaches discussed in this section. However, the focus is on e-business, unlike most of the work on meta-data quality as mentioned earlier. Furthermore, the focus is on assessing the publishing process in order to indirectly improve the QoE and QoBiz. Finally, the work is based on a Business Intelligence (BI) approach, which provides tools for a flexible analysis, including evolution in time in combination with a variety of tools, including statistics, data mining and visualization.

3 Methodology to monitor the quality of content meta-data

In this section, we present a methodology to monitor the quality of meta-data used to describe content in web portals.

3.1 Quality of process in a quality evaluation framework

An e-business company may have success (i.e., achieve a high Quality of Business, QoBiz) depending on whether its services can easily satisfy the needs of its users

(i.e., provide a high Quality of Experience, QoE). The goal is to optimize the quality of the services provided in terms of the following criteria:

Efficiency	how quickly is the answer to a need obtained? Note that we are not concerned with speed in terms of data communication. We are interested in the time required to get to a relevant content, namely in terms of how many steps must the user go through;
Completeness	does the set of content items obtained for a given necessity contain all relevant information? This criterion is related to the notion of recall in information retrieval (Cleverdon et al. 1966; Rijsbergen 1979);
Relevance	does the set of content items obtained for a given necessity include only relevant information? This criterion is related to the notion of precision in information retrieval (Cleverdon et al. 1966; Rijsbergen 1979).

Some of the mechanisms commonly used by web sites to provide their services are search engines and navigation in the site structure. There are many factors affecting the quality of these services: technological structure, sources of content and the processes of publishing and distributing content from the web portal.

In this work, we focus on the publishing process, which consists of the activities of the *authors* and the *editors*. Authors insert content in the web portal, control its access permissions, define the values of meta-data describing it, and integrate it into the organization of the portal (for instance, to classify content in a hierarchy of categories). The editors monitor the publishing process and authorize the publication of content.

This process is very important to the success of a web site. For instance, an inadequate classification of content may make it practically invisible to the user. Nevertheless, it has been ignored in the framework for evaluation of Internet services proposed by Moorsel (2001). We extend this framework by including metrics to assess what we call the *Quality of Process* (QoP). Our goal is to assess the effect of publishing operations on the quality of the services provided by the site.

3.2 Adapted publishing process

In Fig. 1 we describe the standard web publishing process and add two new elements (dashed lines) that are important in our methodology. The standard web publishing process includes editors/authors who insert or change content and respective meta-data. Content is made available on the portal and the user accesses are recorded in the web access logs.

We extend this process with an evaluation tool, called EdMate, which analyzes the quality of the content meta-data (Domingues et al. 2006). Quality is assessed using a number of QoP metrics (Sect. 3.4). The computation of these metrics uses the meta-data describing content, as would be expected, but it may also analyze content and data representing user behavior. The values of those metrics and their evolution in time are represented and summarized in web reports to be used by the Super-Editor and the editors/authors.

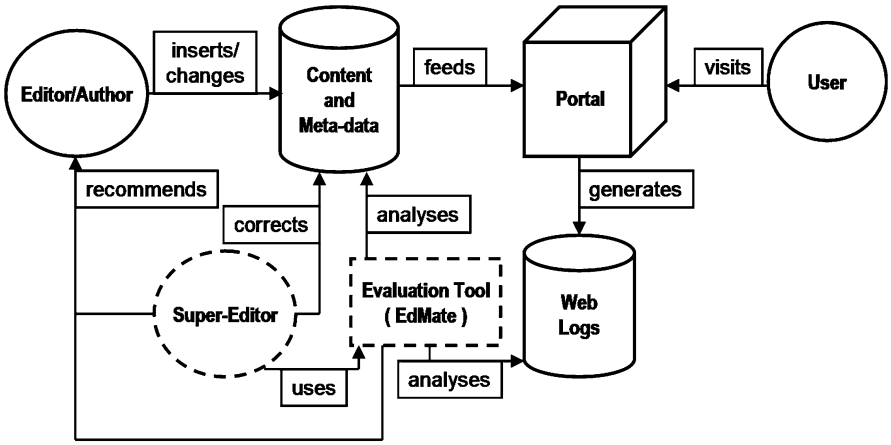

Fig. 1 Publishing process using the proposed methodology

The second new element is the Super-Editor, who processes the output of EdMate and fixes problems directly or makes suggestions to authors. The possible actions for the Super-Editor are:

- Correction of content meta-data;
- Change of the (hierarchical) organization of the contents;
- Change of the schema of meta-database;
- Advice to the authors/editors concerning the publishing process.

3.3 Meta-data quality dimensions

To define a set of QoP metrics, we have adapted the general approach defined by Pipino and Wang (2002), who describes 16 dimensions along which we should assess the quality of data in general. To address the problem of assessing the quality of meta-data in a web portal, we have reinterpreted the 16 dimensions to fit our aims.[2] We have also related those dimensions to the three criteria mentioned earlier, namely efficiency, completeness and relevance. In the following, we reinterpret and comment on each dimension with respect to its application to the quality of meta-data. Furthermore, we have grouped the initial set of dimensions into three large groups, namely *error*, *adequacy* and *value*.

Error: metrics to analyze data problems related mainly with edition lapses. In this group we find, for instance, metrics that show whether a meta-data field has been filled in, and if it contains valid values. The metrics of this group can be classified in one of the following data quality dimensions:

[2] We have ignored the Security dimension, which is not relevant in the publishing process.

- *Believability* Evaluate if the meta-data describe correctly the properties of the content. An incorrect value will lead the user to irrelevant content and miss relevant ones, affecting both completeness and relevance.
- *Completeness* Evaluate if the meta-data describe correctly all the relevant properties of the content. A relevant content may be invisible to a user if it is incompletely described. This dimension is related to the criterion with the same name.
- *Consistent Representation* Evaluate if the meaning of values is always the same across different content. Representing the same concept with different values or associating different concepts with the same value affects both completeness and relevance.
- *Free-of-Error* Evaluate if the meta-data are correct and reliable. Incorrect descriptions will affect completeness.

Adequacy: metrics that are related with the incorrect choice of values to describe content. These metrics can be classified in one of the following data quality dimensions:

- *Accessibility of Content* Evaluate if the meta-data affect the visibility of content. This dimension is related to the efficiency and completeness criteria.
- *Appropriate Amount of Data* Evaluate if the quantity of meta-data is suitable for the users and also if it enables the user to get the adequate amount of content for its needs. An insufficient amount of meta-data affects completeness while too much meta-data and content affect efficiency.
- *Concise Representation* Evaluate if content is described by the smallest set of complete meta-data. A concise description will decrease the amount of information that the user must process, thus increasing efficiency.
- *Ease of Manipulation* Evaluate if the representation of the meta-data is simple enough for the users. A too complex representation will affect efficiency and completeness.
- *Relevancy* Evaluate if the meta-data are informative and useful for the activities of the users. Highly relevant values in the description of content will make the task of identifying relevant content easier, thus improving both efficiency and relevance.
- *Timeliness* Evaluate the extent to which the meta-data enable the user to access relevant content on time. Up-to-date descriptions will increase the probability of finding relevant content and avoiding irrelevant ones. This dimension affects completeness and relevance.

Value: metrics depending on subjective choices of the author, possibly adding value to content and corresponding meta-data. In this group there are metrics that show whether particular meta-data values cause a user to become interested in other content besides the originally intended. The metrics of this group can be classified in one of the following data quality dimensions:

- *Interpretability* Evaluate if the values used by the authors exist in the vocabulary of the users. This dimension affects both completeness and relevance.

- *Objectivity of the Author* Evaluate if the authors describe the content independently of their background. A subjective choice of meta-data values will force the user to try to take the perspective of the author into account, which will affect all three criteria considered.
- *Reputation of the Author* Assign a confidence degree to some value taking into account the quality of meta-data inserted previously by the same author. A more reliable author will need less attention from the Super-Editor. This dimension affects all three criteria considered.
- *Understandability* Evaluate if the semantic of the values is the same for the author and the users. Different understandings affect both completeness and relevance.
- *Value-Added* Assess the utility of the meta-data beyond the primary necessity of the user. A description that leads the user to a relevant content which, additionally, creates a new need is important. This dimension affects completeness and relevance.

As we will discuss in the following section, there are some difficulties in the adaptation and application of these high general principles to the design of concrete objective metrics.

3.4 Meta-data quality metrics

We have designed 31 metrics, covering all the dimensions described above. Table 1 presents a few examples for illustration purposes. The complete list of metrics for measuring the quality of content meta-data is presented in "Appendix".

Many of the data quality concepts, mentioned in the previous section, are quite subjective (e.g., the adequacy of the meta-data values). Furthermore, obtaining explicit satisfaction ratings from the web user is typically difficult. This makes the objective assessment of the quality dimensions described a hard task. Thus, many of the metrics designed only assess the corresponding property indirectly.

To give an example, the *believability* of a descriptor is difficult to assess objectively. However, we can determine the length of the value of a descriptor, such as a keyword. It is expected that a very long keyword is generally less adequate than a shorter one. Therefore, this metric can be used as an estimate of the believability of a descriptor (see *Length of meta-data values* in Table 1). Additionally, a single metric may be used to estimate the quality of a descriptor in terms of more than one dimension. For instance, short search expressions are generally more probable than longer ones. Therefore, the *Length of meta-data values* metric can also be used to assess the *ease of manipulation* dimension. Other example, the *relevancy* of a descriptor is not so easy to assess objectively either. However, we can assess it by determining if the value of a descriptor also appears as a value in another descriptor. For example, if the value of a descriptor appears in the title and/or summary of a content, we can say that the value of the descriptor is relevant for the content (see *Existence in another field* in Table 1). As a final example, the *accessibility of content* is very difficult to be measured. However, we can calculate the number of different values in a descriptor. It is expected that a descriptor with more number

Table 1 Name, dimensions of quality and description of a few metrics

Name:	Length of meta-data values
Dimensions:	Believability, completeness, accessibility of content, concise representation and ease of manipulation
Description:	Number of characters in the value of a meta-data. Extremely large or small values may indicate an inadequate choice of meta-data values to represent the content
Name:	Quantity of meta-data values per content
Dimensions:	Completeness, accessibility of content, appropriate amount of data and concise representation
Description:	Number of different values which are used as meta-data per content. Lower quantities may indicate regular procedures of filling in. Higher quantities may indicate a careful filling in, but maybe with the insertion of values with low description
Name:	Empty meta-data field
Dimensions:	Completeness and accessibility of content
Description:	Number of contents with a given meta-data field not filled in. If a meta-data is used to find a content but the meta-data field is not filled in, the content will not be found
Name:	Existence in another field
Dimensions:	Believability and relevancy
Description:	Meta-data values which appear in another data field. For instance, a meta-data value x may be a good choice if it also appears in the title/summary of a content
Name:	Frequency in search
Dimensions:	Accessibility of content, relevancy, interpretability, understandability and value-added
Description:	Number of meta-data values in the web access logs. For instance, the frequency of a search using a given keyword. If such a keyword is searched for often, probably it will have a high interpretability
Name:	Redundancy of meta-data values
Dimensions:	Concise representation and relevancy
Description:	Conditional probability $P(x\|y)$, where x is one meta-data value of a content, and y is another one. High values may mean that y makes x redundant. This may indicate that implicit practices in the description of content have been developed
Name:	Association between meta-data values
Dimensions:	Concise representation and relevancy
Description:	The confidence level of an association rule $A \rightarrow B$ is an indicator of whether the set of values A makes the set of values B redundant or not. The higher the value, the more redundant B is expected to be. This may indicate that implicit practices in the description of content have been developed

of values is more accessible and vice-versa (see metrics *Quantity of meta-data values per content* and *Empty meta-data field* in Table 1).

The functions used to compute metrics can be based on very simple statistics or more complex methods. For instance, the *Length of meta-data values* metric is computed simply by counting the number of characters in a string, and the *Quantity of meta-data values per content* metric is calculated by counting the number of different values in a meta-data field. The metrics *Empty meta-data field* and *Existence in another field* are computed, respectively, by checking if a meta-data

field does not contain any value and checking if the value of a meta-data is contained in the title/summary of a content. Metrics based on simple frequencies, such as the *Frequency in search* (Table 1), are quite common. Alternatively, metrics can be based on probabilities. The *Redundancy of meta-data values* metric is based on the conditional probability of having a value x, in the description of content, given that a another value y is used (Table 1). An example of the use of a more complex method is given by association rules (Agrawal and Srikant 1994), which are used to compute the *Association between meta-data values* metric (Table 1).

The computation of the metrics is always based on the meta-data. However, in some cases the portal web access log can also be used, such as in the case of the *Frequency in search* metric (Table 1).

4 The EdMate system

Each of the data quality metrics will be computed for a large number of objects. For instance, the *Length of meta-data values* metric mentioned above will be computed for each possible keyword. This yields a huge number of values. Assessing the quality of the meta-data based on all the values computed would be inefficient. Therefore, for the process to be feasible it is necessary to use suitable forms of presentation.

For each metric a small number of *macro indicators* is computed, aggregating the corresponding individual values. The choice of the functional form used to calculate a macro indicator should take into account the semantic of the metric (Pipino and Wang 2002). A suitable function for the *Length of meta-data values* metric example given earlier is the minimum of the frequency values, and the corresponding macro indicator is called *minimum frequency*.

Additionally the Super-Editor should be able to drill-down on the macro indicators in order to obtain more detailed information concerning the values obtained with a metric. For instance, given a low value of the *minimum frequency* macro indicator the Super-Editor may want to find out what the corresponding meta-data descriptor is and which authors have used it. The drill-down mechanism should provide information at different levels of aggregation.

Graphical representation of the values are also used to detect interesting events. For instance, it may be used to provide contextual information, which helps the detection of unusual values. The evolution of the values of the *minimum frequency* macro indicator may show, for instance, that, although the current value is acceptable, the values have been decreasing. This could mean that the authors are describing content less carefully.

4.1 Architecture of the EdMate system

The methodology proposed has been implemented as the EdMate system (Fig. 2). The first version of this system was described in Domingues et al. (2006). It was developed as a set of scripts, which was taking approximately 2 h and 15 min to process around 100 content items and 700 web log accesses, and compute all the

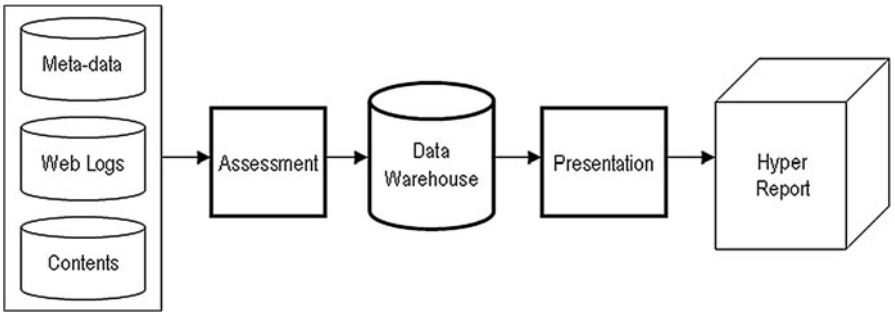

Fig. 2 Architecture of the EdMate system

metrics, statistics and graphics. Therefore, to handle a bigger volume of data, we had to redesign the system. In this section we describe the new version of the EdMate system.

Calculating the whole set of metrics is computationally demanding. To cope with that, we have addressed this issue in two ways. First, the computation of the metrics is done in an incremental way, whenever possible. Second, we have divided the process into two main modules, *Assessment* and *Presentation*. The first module calculates the whole set of metrics, while the second one provides a web interface to analyze the calculated metrics.

4.2 The assessment module

The Assessment module periodically computes the values of the 31 metrics and the macro indicators using data representing the activities of the authors and users, i.e., *Meta-data*, *Web Logs* and *Contents* (Fig. 2). For this new version of the EdMate system, we have adapted the Extraction, Transformation and Loading (ETL) process proposed in Domingues et al. (2007) to pre-process the activity data, compute the quality metrics and load them into a data warehouse. The ETL process is presented in Fig. 3. One of the advantages of this module is that it is developed as a

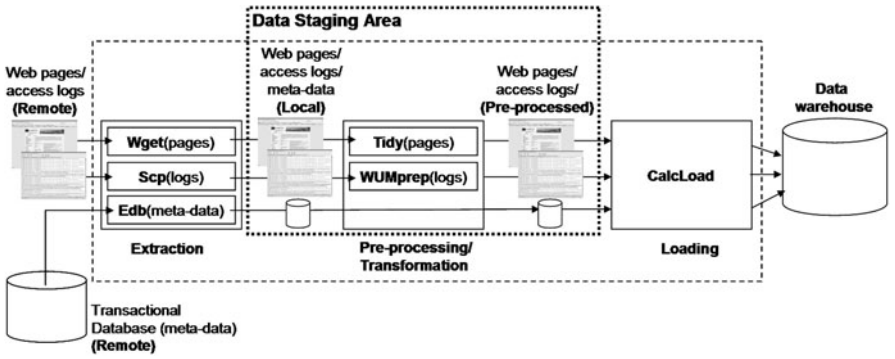

Fig. 3 The extraction, transformation and loading (ETL) process adapted for the EdMate system

composition of different existing tools. Another advantage is that it runs as a batch process, without any manual intervention.

As the name indicates, the process is done in three steps: *extraction*, *pre-processing / transformation* and *loading*. In the *extraction* step, the process creates a local version of (the possibly remote) activity data. This local version is stored in the Data Staging Area (DSA), a simple directory in the file system. For this task, we use ***Wget***,[3] ***Scp***[4] and ***Edb***. *Wget* is a free software for retrieving remote files using HTTP, HTTPS and FTP, which are the most widely used Internet protocols. *Scp* is a software implementing the SCP protocol for secure copying of files between a local and a remote host or between two remote hosts. Finally, *Edb* is a SQL-based component developed by us to select meta-data from a transactional database and create a local version in text files.

In the following step, the local version of the activity data are pre-processed and transformed in useful information ready to compute the quality metrics and be loaded into a data warehouse. For web pages (contents), the process reads the HTML files, and writes clean and well-formed markup in XHTML format. For this task, we use ***Tidy***.[5] This is an open source software and library for checking and generating clean and well-formed XML/XHTML/HTML files. The pre-processing of the access logs consists of merging the log files, removing irrelevant requests and/or data fields, removing robot requests, and identifying users and sessions for the local version of the access logs. We use ***WUMPrep***,[6] a collection of Perl programs supporting data preparation for data mining of web logs. Regarding the meta-data, we do not pre-process and transform them given that our goal is to evaluate their quality.

At this point, we are ready to compute the metrics and load them, together with the activity data, into the data warehouse. Therefore, for the *loading* step, we have implemented a R-based component, called ***CalcLoad***, that computes the quality metrics and load them into the data warehouse. R[7] is an integrated suite of software facilities for data manipulation, calculation and graphical display.

4.3 The data warehouse

In order to provide a suitable repository of data for the EdMate system, we have extended the data warehouse proposed in Domingues et al. (2007) by including tables which make possible the storage of meta-data, metrics and macro indicators (i.e., statistics and graphics). The choice of a data warehouse for the EdMate system instead of a transactional database is motivated by the fact that it is essentially an analytical system. Furthermore, we chose this data warehouse because it has been successfully used to support different web systems (Domingues et al. 2008; Domingues 2008; Carneiro 2008).

[3] http://www.gnu.org/software/wget/.

[4] http://www.openssh.org/.

[5] http://tidy.sourceforge.net/.

[6] http://hypknowsys.sourceforge.net/wiki/Web-Log_Preparation_with_WUMprep/.

[7] http://www.r-project.org/.

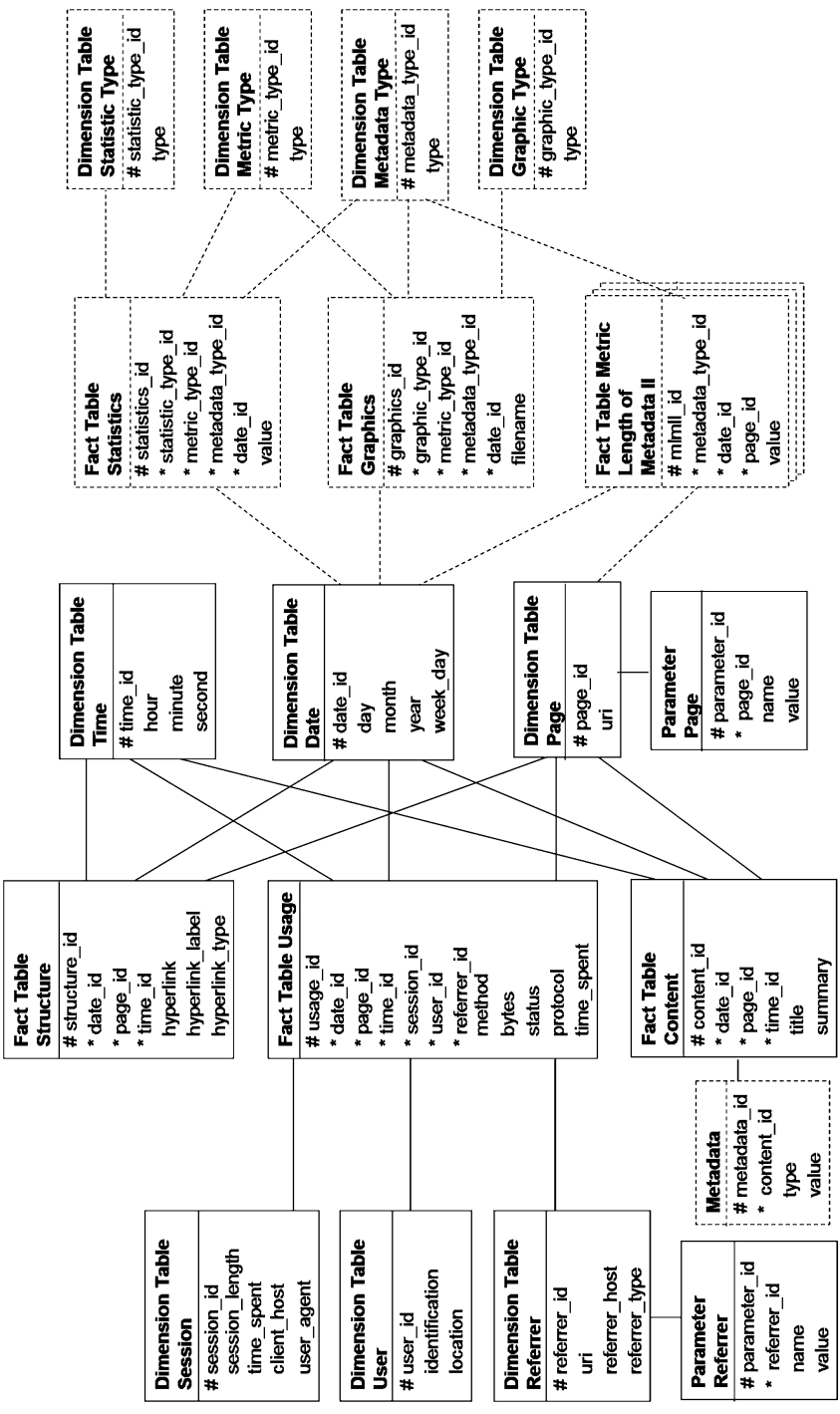

Fig. 4 Star schema of the data warehouse for the EdMate system

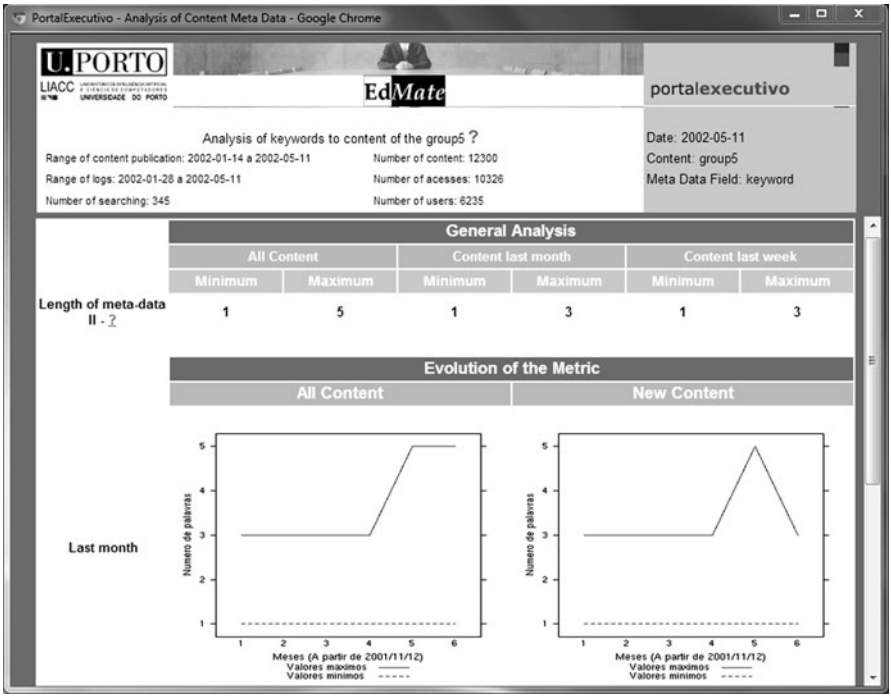

Fig. 5 EdMate screen showing a kind of content meta-data analysis

The extended star schema for the data warehouse is presented in Fig. 4. The original version of the data warehouse only contains fact tables to store data related to structure, usage and content of a web site. The EdMate system provides analyses which are based on meta-data, metrics and macro indicators. Therefore, it is needed to have tables to store these data. In this new version, we have extended the star schema by including such tables. The added tables are represented by the dashed tables in Fig. 4.

In the fact table *Structure*, we store each hyperlink in the web site, and consequently, the hierarchy organization of the site. The information from the web logs are stored in the fact table *Usage*. Finally, we store the content (i.e., title and summary) and its meta-data (i.e., type and value) in the tables *Content* and *Metadata*. In this new version of the data warehouse, the addition of the table *Metadata* and its relationship with the table *Content* allow the storage of different meta-data which belong to a specific content item.

In the star schema, each metric has its own fact table to store its value and information related to it (e.g., type of meta-data assessed by the metric, page which the meta-data are associated to, etc). In Fig. 4, we can see an example of fact table for the metric *Length of Metadata II*, which consists in computing the number of words in a meta-data field. It stores the type of meta-data that is assessed (foreign key *metadata_type_id*), when the metric is calculated (foreign key *date_id*), the web page which the metric is associated to (foreign key *page_id*) and the *value* of the metric.

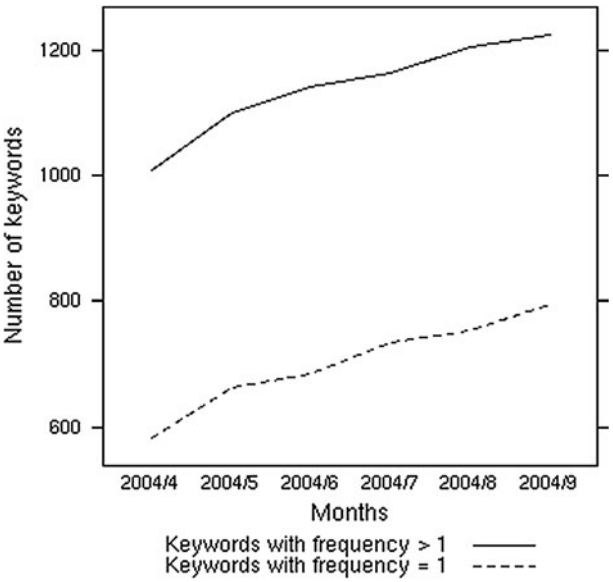

Fig. 6 Evolution of the number of keywords with frequency 1 (*Metric: Singleton meta-data values*)

The statistical indicators and graphics are stored in the fact tables *Statistics* and *Graphics*, which are very close each other in terms of structure (Fig. 4). The fact table *Statistics* stores the type of statistical indicator (foreign key *statistic_type_id*) and the *value* for the statistic. The fact table *Graphics* stores the type of graphical representation (foreign key *graphic_type_id*) and the *file name* for the graphic. Additionally, both tables also store the metric used by the statistics or graphics (foreign key *metric_type_id*), the type of meta-data assessed by the metric (foreign key *metadata_type_id*) and the date of computation (foreign key *date_id*). The types of statistical indicators, metrics, meta-data and graphics are stored, respectively, in the dimension tables *Statistic Type*, *Metric Type*, *Metadata Type* and *Graphic Type*.

The new fact tables added to the data warehouse facilitate the computation of the different analyses provide by the EdMate system, as for example, the OnLine Analytical Processing (OLAP) analyses (Malinowski and Zimnyi 2008). We have implemented the data warehouse using the Oracle Database[8] to take advantage of its performance enhancement features: partitioned tables, bitmap index and materialized views.

4.4 The presentation module

The Presentation is a module developed in PHP[9] that accesses the metrics, statistical indicators and graphics, stored in the data warehouse, to generate the *Hyper Report*, which is accessed using a web browser. We have decided to implement this module

[8] http://www.oracle.com/us/products/database/.

[9] http://www.php.net/.

as a web based system, that is more manageable and deployable, reducing cost of development and facilitating the access of end users. In Fig. 5, we present a screen of the EdMate system showing the metric *Length of meta-data II*. At the top, we have some information about the data which we are analyzing, such as number of content items, accesses and users, range of the logs, and so forth. In the middle, we can see the statistical indicators of the metric: minimum and maximum values. Finally, at the bottom, we can see the evolution of the metric in graphical representations.

We can explore the metrics from different angles using OLAP analyses (Malinowski and Zimnyi 2008). For instance, if the global value of the metric *Length of meta-data II* is very large, we may have a more detailed view, e.g., by analyzing its values aggregated by day (drill down operation).

5 Case study

In this section we describe the application of the proposed methodology to PortalExecutivo (PE), a Portuguese web portal targeted to business executives. The business model is subscription-based, which means that only paying users have full access to content through web login. However some content is freely available and users can freely browse the site's structure. Content is provided not only by PE but also by a large number of partners. The goal of PE is to facilitate the access of its members to relevant content. Value is added to the contributed content by structuring and interrelating them. This is achieved by filling in a rich set of meta-data fields, including keywords, categories, relevant companies, source, authors, among others. Therefore, the problem of meta-data quality is essential for PE.

5.1 Results

Here we illustrate the kind of analysis that can be made using the EdMate system. We also demonstrate that the publishing process can be changed based on its results so that the quality of the meta-data is improved.

Since the results of queries to the search engine are affected by the quality of the keywords used to describe content, we focus on this meta-data field. The meta-data used is relative to the period April/September 2004.[10]

Concerning the quality of meta-data, Fig. 6 shows that the number of keywords which is used only once is very high (i.e., around 40 % of the total of keywords). On the one hand, some of these are caused by typographical errors, which means that this metric can be associated with the *free-of-error* dimension. On the other, this value indicates that the potential of keywords to interrelate content from different sources is not being adequately exploited (*relevancy* and *value-added* dimensions). Therefore, the keywords meta-data field can not be an information integrator. The analysis indicates the existence of typographic errors, or the necessity of increasing the sharing of the values of the keywords meta-data field among the contents. Since this analysis, the company has already adopted techniques for higher keyword sharing and verification, such as the use of a keyword dictionary.

[10] We only have the authorization to publish results for this period of time.

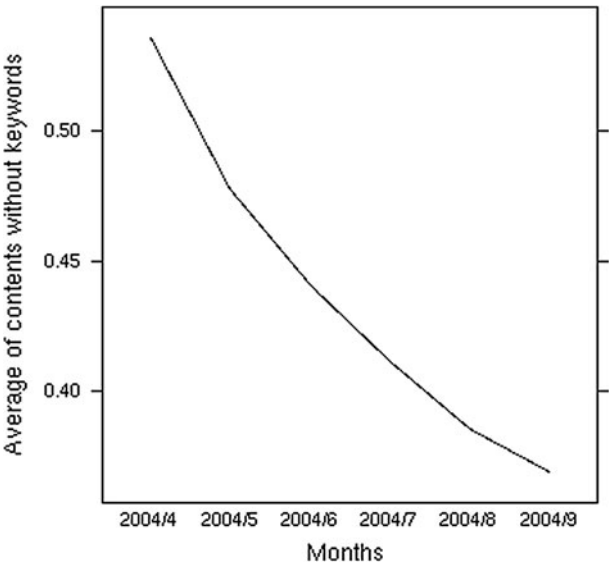

Fig. 7 Evolution of the number of keywords not filled in (*Metric: Empty meta-data field*)

The results obtained with EdMate are not only useful to detect problems with data quality but also to trigger corrective actions and monitor them. Figure 7 shows that in April more than 50 % of content did not have any keywords filled-in (*completeness* dimension).

This was noticed by the Super-Editor at that time and, consequently a semi-automatic procedure to support the process of filling-in keywords was implemented (*completeness* dimension). The same figure shows that this procedure has brought a significant improvement to the quality of meta-data, with a steady decrease of this metric down to less than 40 % in September.

This kind of analysis also enables the Super-Editor to keep an up-to-date perspective on the publishing process. Although the number of contents without keywords has decreased, Fig. 8 shows that, in September, a great quantity of contents (i.e., more than 30 % of the contents) does not have the keywords meta-data field filled in, and very few content items have more than 4 keywords. The Super-Editor may find this insufficient and, thus instruct the authors accordingly. Additionally, Fig. 9 shows that the maximum number of keywords associated with a content item is 17. This may look suspicious to the Super-Editor, who may identify the corresponding item and correct its description, if necessary.

Once the contents are retrieved by using the keywords through search engine, it would be important that such keywords appear in the title or the description (summary) of the content. Figures 10 and 11 show the evolution of the number of keywords in the title and description of the contents. In Fig. 10, the Super-Editor can see that there are much more keywords out of the title than in it, and that appropriate actions must to be taken, as for example, notifying the authors to use keywords that appear in the title of the content. In Fig. 11, although there are much

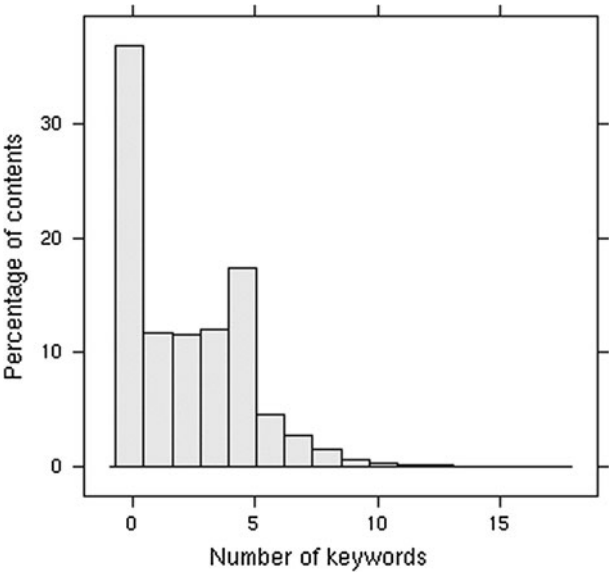

Fig. 8 Histogram of the number of keywords by content (*Metric: Quantity of meta-data values per content*)

more keywords in the description of the contents, it shows that since May less keywords have appeared in the description of the content . The Super-Editor may see this and, thus instruct the authors accordingly.

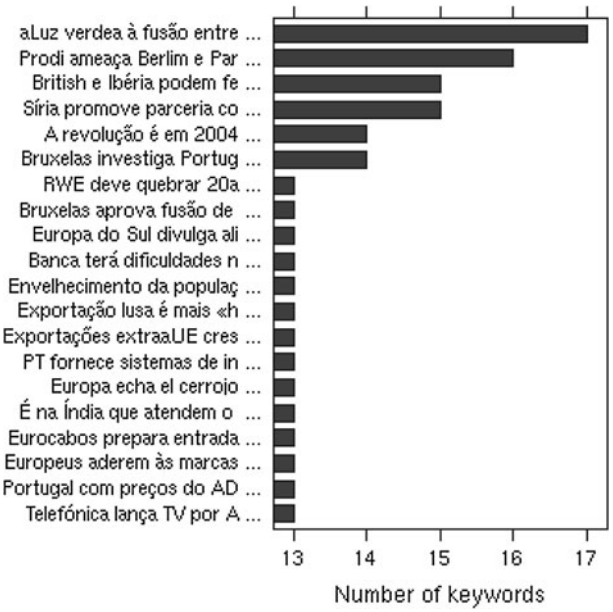

Fig. 9 Top 20 of contents with the largest numbers of keywords (*Metric: Quantity of meta-data values per content*)

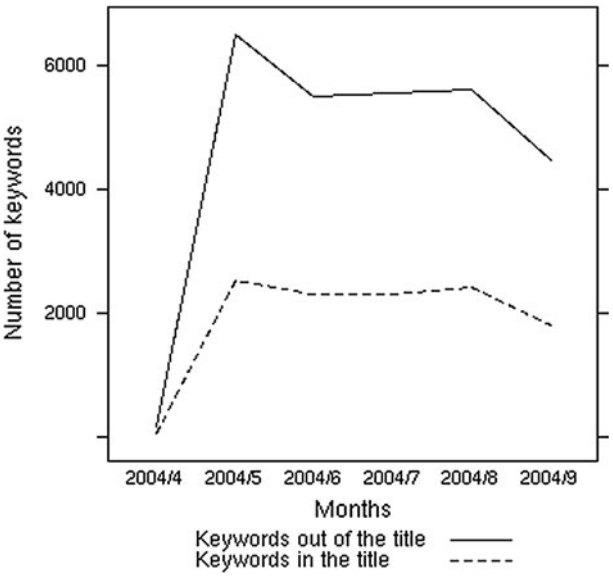

Fig. 10 Evolution of the number of keywords in the title (*Metric: Existence in another field*)

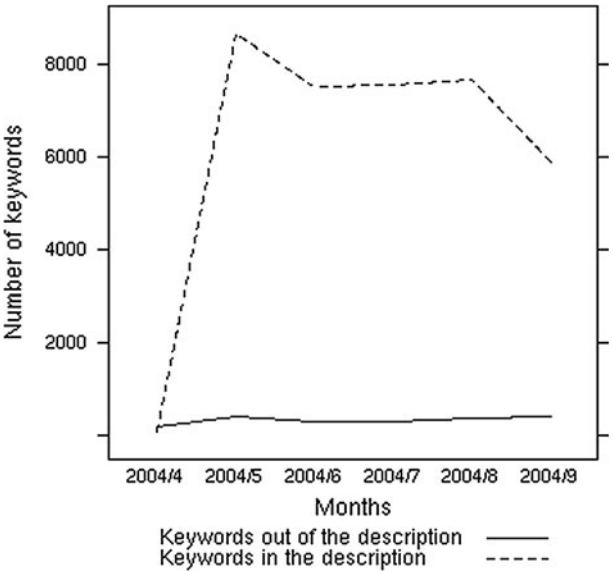

Fig. 11 Evolution of the number of keywords in the description (summary) (*Metric: Existence in another field*)

In another metric, we have used the confidence of association rules to determine keywords more frequently used together. Additionally, we can provide a graphical representation of the associations between keywords (Fig. 12). We observed that

Fig. 12 Relationships between keywords obtained using association rules (*Metric: Association between meta-data values*)

often a general keyword (e.g., fiscality—*fiscalidade*) is associated with a more specific one (e.g., international taxation—*tributação internacional*). This implicit structure of the keywords, unveiled by the discovered association rules, enables the detection of incorrect descriptions (*believability*, *concise representation* and *reputation of the author* dimensions).

5.2 Lessons learnt

These results show that the proposed methodology enables:

- an assessment of the quality of the meta-data, triggering corrective action;
- monitoring of corrective procedures;
- an up-to-date perspective on the publishing process.

In summary, after the analysis of the metrics, the Super-Editor can take some actions to improve the quality of contents meta-data and consequently increase the quality of the web portal. In fact, this has already happened in our application: errors in the content descriptors have been detected and the Super-Editor has taken concrete actions to change the data and the manual content description process. In the future, the actions to improve the quality of the web portal can be automated using artificial intelligence and data mining techniques.

6 Conclusions and future work

Many web portals have a distributed model for the contribution of content. No matter how strict the publishing process is, low quality meta-data will sometimes be used to describe content. This decreases the quality of the services provided to the users.

In this paper we presented a methodology to monitor the quality of meta-data used to describe content in web portals. We also described a general architecture for a system to support the proposed methodology. We have successfully applied the methodology to a portal for business executives. Besides enabling the assessment of the quality of the meta-data, it enables the monitoring of corrective actions and it provides an up-to-date perspective of the publishing process.

As future work, we plan to formalize the quality dimensions (Sect. 3.3) in order to support a more systematic process of designing new metrics. We also plan to apply other statistical and data mining techniques to improve the quality assessment process. For instance, clustering methods (Velasquez and Palade 2008) can be used to obtain groups of authors with similar behaviors in terms of data quality. This not only enables a different perspective on their publishing process but also different corrective actions can then be taken upon different groups.

Additionally, those techniques can be used to extend the quality assessment process with tools to support both authors and the Super-Editor to fill-in and correct meta-data. For instance, classification methods (Velasquez and Palade 2008) can be used to suggest keywords depending on the content.

Finally, to enhance the applicability of the methodology proposed, we need to integrate the Quality of Process metrics into a more general quality assessment framework, including other levels, such as Quality of Service, Quality of Experience and Quality of Business (Moorsel 2001).

Acknowledgments This work was partially funded by PortalExecutivo. The authors are grateful to PortalExecutivo for their support, and, in particular, to Rui Brandão and Carlos Sampaio for their collaboration.

Appendix: Metrics for measuring the quality of content meta-data

See Table 2.

Table 2 Name, dimensions of quality and description of metrics for measuring the quality of content meta-data

Name:	Association between meta-data values
Dimensions:	Concise representation and relevancy
Description:	The confidence level of an association rule $A \rightarrow B$ is an indicator of whether the set of values A makes the set of values B redundant or not. The higher the value, the more redundant B is expected to be. This may indicate that implicit practices in the description of content have been developed
Name:	Redundancy of meta-data values
Dimensions:	Concise representation and relevancy
Description:	Conditional probability $P(x\|y)$, where x is one meta-data value of a content, and y is another one. High values may mean that y makes x redundant. This may indicate that implicit practices in the description of content have been developed
Name:	Number of shadow contents
Dimensions:	Believability, concise representation and relevancy

Table 2 continued

Description:	A content c_1 is shadow of a content c_2 if the set of meta-data values in c_2 is a super-set of c_1. The access to a content that is shadow of other contents may be very difficult using the meta-data
Name:	Length of meta-data I
Dimensions:	Believability, completeness, accessibility of content, concise representation and ease of manipulation
Description:	Number of characters in a meta-data field. Extremely large or small values may indicate an inadequate choice of meta-data to represent the content
Name:	Length of meta-data II
Dimensions:	Believability, completeness, accessibility of content, concise representation and ease of manipulation
Description:	Number of words in a meta-data field. Extremely large or small values may indicate an inadequate choice of meta-data to represent the content
Name:	Length of title/summary I
Dimensions:	Believability, completeness, accessibility of content, concise representation and ease of manipulation
Description:	Number of characters in the title/summary. Large or small values mean that the choice of the title or summary may not be adequate
Name:	Length of title/summary II
Dimensions:	Believability, completeness, accessibility of content, concise representation and ease of manipulation
Description:	Number of words in the title/summary. Large or small values mean that the choice of the title or summary may not be adequate
Name:	Length of title/summary III
Dimensions:	Believability, completeness, accessibility of content, concise representation and ease of manipulation
Description:	Number of phrases in the title/summary. Large or small values mean that the choice of the title or summary may not be adequate
Name:	Extreme frequency of meta-data values
Dimensions:	Believability and appropriate amount of data
Description:	Number of contents which contain a given meta-data value. High values may indicate that the number of contents selected by using such a meta-data will be very high
Name:	Extreme frequency of meta-data values, by editor/author
Dimensions:	Believability, appropriate amount of data and reputation of the author
Description:	Number of contents, grouped by editor/author, which contain a given meta-data value
Name:	Frequency in search
Dimensions:	Accessibility of content, relevancy, interpretability, understandability and value-added
Description:	Number of meta-data values in the web access logs. For instance, the frequency of a search using a given keyword. If such a keyword is searched for often, probably it will have a high interpretability
Name:	Shared meta-data values
Dimensions:	Believability, relevancy and value-added
Description:	Number of meta-data values which are used by at least two different contents. The shared values allow the relationship among contents
Name:	Degree of sharing among editors/authors
Dimensions:	Objectivity of the author

Table 2 continued

Description:	Number of different editors/authors who use/share a same meta-data value
Name:	Empty meta-data field
Dimensions:	Completeness and accessibility of content
Description:	Number of contents with a given meta-data field not filled in. If a meta-data is used to find a content but the meta-data field is not filled in, the content will not be found
Name:	Empty meta-data field, by editor/author
Dimensions:	Completeness, accessibility of content and reputation of the author
Description:	Number of contents, grouped by editors/authors, with a given meta-data field not filled in
Name:	All empty meta-data fields
Dimensions:	Completeness and accessibility of content
Description:	Number of contents with all meta-data fields not filled in. Contents without any meta-data may indicate errors of publication
Name:	All empty meta-data fields, by editor/author
Dimensions:	Completeness, accessibility of content and reputation of the author
Description:	Number of contents, grouped by editors/authors, with all meta-data fields not filled in. Here, we can evaluate the not filling in of meta-data by editors/authors
Name:	Length of meta-data values
Dimensions:	Believability, completeness, accessibility of content, concise representation and ease of manipulation
Description:	Number of characters in the value of a meta-data. Extremely large or small values may indicate an inadequate choice of meta-data values to represent the content
Name:	Quantity of meta-data values
Dimensions:	Completeness
Description:	Number of different values which are used as meta-data
Name:	Quantity of meta-data values per content
Dimensions:	Completeness, accessibility of content, appropriate amount of data and concise representation
Description:	Number of different values which are used as meta-data per content. Lower quantities may indicate regular procedures of filling in. Higher quantities may indicate a careful filling in, but maybe with the insertion of values with low description
Name:	Singleton meta-data values
Dimensions:	Free-of-error
Description:	Meta-data values which are used only once. This metric may indicate typographical errors
Name:	Repeated meta-data values
Dimensions:	Concise representation and free-of-error
Description:	Number of repeated meta-data values in a content
Name:	Contents with repetition
Dimensions:	Believability, consistent representation and free-of-error
Description:	Number of contents with the same meta-data values
Name:	Existence in another field
Dimensions:	Believability and relevancy
Description:	Meta-data values which appear in another data field. For instance, a meta-data value x may be a good choice if it also appears in the title/summary of a content
Name:	Isolated usage
Dimensions:	Believability, objectivity of the author and reputation of the author

Table 2 continued

Description:	Number of editors/authors who always use the same meta-data values
Name:	Latency of free access
Dimensions:	Believability and timeliness
Description:	If the free access date for a content is set to a date that comes quite later the publication date (for instance, more than 1 year), the value for this meta-data may not be adequate
Name:	Invalid free access range
Dimensions:	Free-of-error
Description:	he free access range is invalid when the start date for free access is set to a date that comes after the end date for free access
Name:	Shorted free access range
Dimensions:	Accessibility of content
Description:	This metric indicates whether there is a small difference between the start and the end date for the free access or not
Name:	Extreme price
Dimensions:	Believability
Description:	Higher or lower selling prices for a content may indicate an inadequate choice for this meta-data
Name:	Invalid price
Dimensions:	Free-of-error
Description:	The selling price of a content is invalid when it is negative
Name:	Depth of the hierarchy
Dimensions:	Accessibility of content and ease of manipulation
Description:	This metric shows the depth of the hierarchy of a web site. A great number of levels in the hierarchy means that will be difficult to find a content if it is in the lowest levels of this hierarchy

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of 20th international conference on very large data bases, pp 487–499. http://citeseer.nj.nec.com/agrawal94fast.html. Accessed 30 Nov 2009

Baeza-Yates R, Rello L (2012) On measuring the lexical quality of the web. In: Proceedings of the 2012 Joint WICOW/AIRWeb workshop on web quality (WebQuality 2012), pp 1–6. doi:10.1145/2184305.2184307

Berendt B (2002) Using site semantics to analyze, visualize, and support navigation. Data Min Knowl Discov 6(1):37–59. doi:10.1023/A:1013280719795

Blanco L, Crescenzi V, Merialdo P, Papotti P (2011) Characterizing the uncertainty of web data: models and experiences. In: Proceedings of the 2011 joint WICOW/AIRWeb workshop on web quality (WebQuality 2011), pp 1–8. doi:10.1145/1964114.1964116

Bruce TR, Hillmann D (2004) The continuum of metadata quality: defining, expressing, exploiting. American Library Association, Chicago

Cadez I, Heckerman D, Meek C, Smyth P, White S (2003) Model-based clustering and visualization of navigation patterns on a web site. Data Min Knowl Discov 7(4):399–424. doi:10.1023/A:1024992613384

Carneiro A (2008) Using web data for measuring the effectiveness of an e-commerce site. Master's thesis, University of Porto, Faculty of Economics, Portugal

Cleverdon CW, Mills J, Keen M (1966) Aslib cranfield research project—factors determining the performance of indexing systems; volume 1, design; part 1, text. Tech. rep., Cranfield University. http://hdl.handle.net/1826/861. Accessed 30 Nov 2009

Das R, Turkoglu I (2009) Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. Exp Syst Appl Int J 36:6635–6644. doi:10.1016/j.eswa.2008.08.067

Domingues MA (2008) An independent platform for the monitoring, analysis and adaptation of web sites. In: Pu P, Bridge DG, Mobasher B, Ricci F (eds) Proceedings of the 2008 ACM conference on recommender systems, RecSys 2008, Lausanne, Switzerland, October 23–25, 2008, pp 299–302

Domingues MA, Soares C, Jorge AM (2006) A web-based system to monitor the quality of meta-data in web portals. In: WI-IATW '06: proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, IEEE Computer Society, Hong-Kong, China, pp 188–191. doi:10.1109/WI-IATW.2006.24

Domingues MA, Jorge AM, Soares C, Leal JP, Machado P (2007) A data warehouse for web intelligence. In: Proceedings of the 13th Portuguese conference on artificial intelligence, pp 487–499

Domingues MA, Jorge AM, Soares C, Leal JP, Machado P (2008) A platform to support web site adaptation and monitoring of its effects: a case study. In: Proceedings of the 6th workshop on intelligent techniques for web personalization and recommender systems (ITWP 2008), Chicago, Illinois, pp 29–36

Fluit C, Wester J (2002) Using visualization for information management tasks. In: International conference on information visualisation

Guy M, Powell A, Day M (2004) Improving the quality of metadata in eprint archives. Ariadne (38). http://www.ariadne.ac.uk/issue38/guy/. Accessed 11 Sept 2012

Isinkaye FO, Robert ABC, Ojokoh BA (2012) An evaluation of metadata integrity in textual documents. J Libr Metadata 12(1):1–14. doi:10.1080/19386389.2012.652565

Lex E, Voelske M, Errecalde M, Ferretti E, Cagnina L, Horn C, Stein B, Granitzer M (2012) Measuring the quality of web content using factual information. In: Proceedings of the 2012 joint WICOW/AIRWeb workshop on web quality (WebQuality 2012), pp 7–10. doi:10.1145/2184305.2184308

Malinowski E, Zimnyi E (2008) Advanced data warehouse design: from conventional to spatial and temporal applications (Data-Centric Systems and Applications). Springer Publishing Company, Incorporated

Moorsel AV (2001) Metrics for the internet age: quality of experience and quality of business, fifth performability workshop. Tech. rep., Software Technology Laboratory—HP Laboratories Palo Alto. http://www.hpl.hp.com/techreports/2001/HPL-2001-179.pdf. Accessed 20 Nov 2006

Nichols DM, Chan CH, Bainbridge D, McKay D, Twidale MB (2008) A lightweight metadata quality tool. In: Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries (JCDL 2008), pp 385–388. doi:10.1145/1378889.1378957

Ochoa X, Duval E (2006) Towards automatic evaluation of learning object metadata quality. In: Proceedings of the 2006 international conference on advances in conceptual modeling: theory and practice. Springer, Berlin, Heidelberg, pp 372–381. doi:10.1007/11908883_44

Ochoa X, Duval E (2009) Automatic evaluation of metadata quality in digital repositories. Int J Digit Libr 10(2–3):67–91. doi:10.1007/s00799-009-0054-4

Park JR (2009) Metadata quality in digital repositories: a survey of the current state of the art. Catalog Class Q 47(3–4):213–228. doi:10.1080/01639370902737240

Pipino L L, Lee YW, Wang RY (2002) Data quality assessment. Commun ACM 45(4):211–218

Rijsbergen CJV (1979) Information retrieval. Butterworth-Heinemann, Newton, MA, USA

Soares C, Jorge AM, Domingues MA (2005) Monitoring the quality of meta-data in web portals using statistics, visualization and data mining. In: Proceedings of Twelfth Portuguese conference on artificial intelligence (EPIA 2005), LNAI 3808, Covilhã, Portugal, pp 371–382

Spiliopoulou M, Pohle C (2001) Data mining for measuring and improving the success of web sites. Data Min Knowl Discov 5(1–2):85–114

Stvilia B, Gasser L, Twidale MB, Shreeves SL, Cole TW (2004) Metadata quality for federated collections. In: 9th international conference on information quality (IQ 2004), pp 111–125

Velasquez JD, Palade V (2008) Adaptive web sites: a knowledge extraction from web data approach—volume 170 frontiers in artificial intelligence and applications. IOS Press, Amsterdam, The Netherlands

Vuong BQ, Lim EP, Sun A, Chang CH, Chatterjea K, Goh DHL, Theng YL, Zhang J (2007) Key element-context model: an approach to efficient web metadata maintenance. In: ECDL'07:

Proceedings of the 11th European conference on research and advanced technology for digital libraries, Springer, Berlin, Heidelberg, pp 63–74. doi:10.1007/978-3-540-74851-9_6

Zaïane OR, Xin M, Han J (1998) Discovering web access patterns and trends by applying olap and data mining technology on web logs. In: Proceedings of the advances in digital libraries conference (ADL-1998), IEEE Computer Society, Washington, DC, USA, pp 19–29