PRESERVING WEB SITES: A DATA QUALITY APPROACH

(Complete Paper)

Cinzia Cappiello
Chiara Francalanci
Barbara Pernici
Politecnico di Milano, Milano, Italy
{cappiell, francala, pernici}@elet.polimi.it

Abstract: The term *preservation* indicates the ability to prevent losses of information as a consequence of updates, by storing all significant versions of web data and, thus, creating a web library. This paper proposes a methodology to support the selection of relevant information to be preserved. Data quality evaluation techniques are applied to define and measure the relevance of web information. The evaluation of data quality is performed through a software architecture, named Quality Factory, which is responsible for monitoring and assessing the quality of web pages both at creation and update time. Experimentation is conducted on the web site of Politecnico di Milano Technical University. Results show how the percentage of useful information that can be lost as a consequence of updates is about 50% of the site's pages each year. The methodology can obviate this problem and, at the same time, avoid the indiscriminate preservation of low-quality data.

Key Words: Data quality, Preservation, Web data

1. INTRODUCTION

The number of documents that are managed in web format is constantly growing. Even if they are printed and stored in paper format at some time in their life cycle, web documents can undergo several electronic updates. Contrary to paper updates, electronic updates typically replace previous web documents and can generate critical losses of information. The term *preservation* indicates the ability to prevent these losses of information, by storing all significant versions of web documents and, thus, creating a web library.

The research questions related to preservation are extensive and complex. For example, what is the information to be preserved? How should it be preserved? And, finally, who will be responsible for preservation? This paper focuses on the first question that is the selection of relevant information to be preserved. Currently, there is no standard and comprehensive methodology that helps evaluating the relevance of information. However, past research on preservation provides an indication of the quantitative and qualitative parameters for the evaluation of relevance. Quantitative parameters are typically related to the space limits of physical storage devices and measure various proxies of the storage cost associated with a web document [15]. Qualitative parameters are used to determine whether a document should be preserved and, if so, what are the benefits from its preservation. In general, relevance grows with benefits and decreases with storage costs.

This paper proposes a methodology to support the preservation process over the entire life cycle of information, from creation, to publication, access, classification and storage. The methodology is based on the assumption that who creates information is also responsible for its preservation. This reduces the impact of storage costs, as the web library is distributed and costs are shared among multiple parties. Accordingly, the methodology focuses on qualitative parameters. Past research on data quality provides a

consolidated set of quality dimensions and corresponding operating measures which can support the evaluation of relevance. Data quality evaluation techniques are also suitable for automation and, thus, can be implemented as semi-automatic software procedures that minimize human effort. They also provide a contextual assessment of preservation benefits, as they originate from the generally accepted definition of quality as "fitness for use" [18][24].

The presentation is organized as follows. Section 2 reviews the literature on preservation. In Section 3, data quality evaluation techniques are discussed. The preservation methodology is introduced in Section 4, by describing methodological phases, required resources and operating procedures. The evaluation of data quality is performed through a software architecture, named Quality Factory, which is presented in Section 5. Finally, Section 6 reports experimental results.

2. THE PRESERVATION OF WEB DOCUMENTS: STATE OF THE ART

A common web maintenance practice is to update pages by overwriting previous versions of information. Linking the updated versions of a page to the previous URL avoids the revision of the site's structure. Furthermore, preserving an online copy of all versions of documents would considerably increase the size of the memory needed to store the site, thus raising web management costs. Multiple online versions of the same page may also be confusing to users accessing the site. On the other hand, the Internet is continuously developing and the mass of information that is lost with updates grows accordingly. It has been demonstrated that the average lifetime of a web page is about 100 days and, overall, the web changes completely about four times in a year [4].

Classical techniques for information preservation are unsuitable for the electronic versions of documents. For example, all techniques employed in traditional libraries for the preservation of paper documents do not apply to web libraries. Each time that a traditional library receives a new book, the catalogue is updated by assigning a unique inventory number and a press-mark to the new book. Two labels are applied on the book: an identifier that specifies the library, the inventory number and the press-mark and a second label that indicates the date when the borrower must return the book. Finally, the book is stamped and covered with a transparent material (Digifix) in order to protect the identifying label. Different rules are followed to catalogue a new cd-rom. The identifiers, inventory and press-mark, are written on the upper side of the disk with permanent ink and an identifying label is applied on the case. No tutelage is operated against the violation of the information on the disk. The bar code can only be applied on the case, giving rise to errors due to accidental exchanges of cases between cd-roms. Clearly, these preservation techniques cannot be applied to documents that are devoid of physical support, such as web pages, and must be entirely redesigned.

Since the introduction of the web as a global information resource, several organizations have attempted its preservation. Initiatives have taken two different approaches [13]. The first approach suggests the selection of a subset of web pages to preserve; the second uses automatic programs to collect and preserve web sites indiscriminately. In most cases, selection criteria consist of a set of guidelines helping the preservation expert to make selection decisions. Quantitative measures are seldom used to make preservation decisions and, in particular, data quality parameters have not been previously applied.

An important research project on the preservation of web documents following the first approach is PANDORA, launched by the National Library of Australia (NLA), which, since 1997, has been preserving all online documents that concern Australian culture [17]. Recently, NLA has been able to create a logo, which identifies the web sites that have joined the project and are preserved. By February 2003, about 15 million files had been preserved [17]. The huge amount of electronic publications and the low value of many online documents have raised the need for a selection model. Within PANDORA, a document is judged suitable for national preservation if most of its content is about Australia or about a topic of great social, political, cultural, scientific or economic importance for the nation, or if the author is Australian. Relevant topics have been classified according to guidelines established by NLA. If a

document is available in both paper and electronic format, the National Library will prefer the paper version. The online version is selected only if it contains value-added information compared to the paper version. The frequency of preservation varies according to the characteristics of individual web site. For example, monographs need to be preserved only once, while periodicals are archived on a weekly basis. If the web site is large, only a subset of pages is preserved and, in any case, links to external pages are not followed. The quality of web pages is not evaluated.

A similar approach is followed by the "Britain on the Web" project, conducted by the British Library from 2001 to 2002 [10]. One hundred web sites have been selected as a snapshot of the UK web activity to be preserved.

A more complex project by the French national library has investigated the preservation of the French web as part of its responsibilities [3]. The project started from experiments on collecting web sites conducted by the Biblioteque nationale de France (BNL) in association with INRIA (the French National Institute for Research in Computer Science and Automatic Control). Based on these initial experiments, the research team has identified a need for automatic support in collecting web pages. The method selects sites from the "French web" according to language and location criteria. An architecture, Xyleme, has been developed to retrieve and store HTML and XML pages [2]. Pages can be stored on local devices or transferred to a remote storage location through FTP. An innovative feature of this architecture is the ability to automatically update a selected set of pages with a refresh period that can either be set by users or automatically calculated on the basis of the frequency of updates. 34 site owners have joined the initiative. Although this project does not explicitly refer to data quality criteria, typical data quality dimensions are used de facto, such as update frequency as a proxy of volatility [11].

A project following the non-selective preservation approach has been initiated in 1996 by the Internet Archive, which periodically "takes a snapshot" of the net and creates a worldwide backup of all sites. This project has the most ambitious goals in terms of the number of web pages to be preserved. Pages are collected and then analyzed by a commercial company, called Alexa Internet, to build web statistics. Password-protected pages and information created dynamically are not preserved. Similar to PANDORA, the quality of web information is not evaluated, while selection is based on predefined and context-independent criteria.

From an academic perspective, there are several projects on preservation research issues. The InterPARES Project (International Research on Permanent Authentic Records in Electronic Systems) [16] is an important international research project involving archivists, computer science engineers, and industries. The first goal of this project is to develop the theoretical and methodological knowledge required for long-term preservation of digital information. InterPARES is coordinated by the School of Library, Archival and Information Studies of the University of British Columbia. The first result has been the identification of appropriate metadata for the authentication of web documents. Preservation issues will be analyzed during the second phase of the project between 2003 and 2006.

Another academic project is ERPANET, whose main goal is to establish a European Initiative to serve as a knowledge-base in the area of preservation of cultural heritage and scientific digital objects [14]. Indeed, ERPANET promotes knowledge exchanges among individuals and institutions on state-of-the-art developments in digital preservation. Many organizations from the ICT, entertainment and media (e.g. broadcasting) industries participate in the project by providing multidisciplinary knowledge and shared resources. More specifically, the ERPANET team offers a range of services, such as content creation, advisory, training and workshops. However, ERPANET does not directly conduct research to develop new preservation methods and tools, but provides the infrastructure for proactive cooperation and dissemination of research results and experiences in the preservation of digital objects.

In the next section, data quality techniques are reviewed and their application to the selection process of preservation is discussed.

3. DATA QUALITY DIMENSIONS

Data quality has been successfully applied within database management processes to eliminate low-quality data from organizational databases. This paper attempts the extension of the positive data quality experience to selection decisions in web preservation procedures. As a selection criterion, data quality provides a consolidated set of evaluation dimensions to make a distinction between high and low quality information. This distinction can be made automatically, while quality thresholds can be selected by experts based on contextual criteria. Several authors define the quality of data as their "fitness for use", i.e. the ability of a data collection to meet users' requirements [24][18]. The notion of quality is therefore tightly related to the applications managing data and to the portion of the real world modeled by such data. The data quality literature provides the following consolidated classification of data quality dimensions:

- Relevance, granularity and level of detail, which are associated with data views.
- Accuracy, consistency, currency and completeness, which are associated with data values.
- Format and ease of interpretation, which are associated with the presentation of data.
- Privacy, security, and ownership, which are general dimensions.

Preservation is concerned with the history of data instances that is data values at different time points. Accordingly, this paper focuses on quality dimensions associated with data values, namely *consistency, accuracy, completeness,* and *currency*. In the following, the definition of these dimensions in the data quality literature is reviewed and their application to the evaluation of web documents is discussed.

Data *consistency* is usually related to general aspects of data. In particular, data values, logical and physical representations are considered. In [21], consistency is defined from three different perspectives: view consistency, value consistency and representation consistency. View consistency is further distinguished into semantic and structural consistency. Data are semantically consistent if the definition of entities resolves issues involving their relationship. Instead, data are structurally consistent if attributes have the same structure for all entity types. Value consistency involves that data have the same value in all their representations. In [21], data consistency requires that data values are not in contrast with each other and attribute consistency holds for each value domain, either discrete or continuous. Representation consistency indicates that physical instances of data comply with their formats. From an evaluation perspective, consistency is a result of a verification effort that measures how well different data items agree with each other or with specified criteria [21]. Specifically, data are consistent with respect to a set of constraints if they satisfy all constraints in the set. As regards representation consistency, constraints force the compliance with a standard format.

Representation consistency is most useful to evaluate the consistency of a web page. A web page is typically composed by multiple sections designed to show a specific object. It is possible to evaluate the correspondence between the format of an object that is placed in a given section and the expected format. By conducting this evaluation for all objects, an aggregate measure of consistency can be obtained as a weighed sum of consistency values for all objects. This aggregate measure of consistency can be associated with either a page, a set of pages or a section. In turn, aggregate measures can be combined with the consistency values associated with individual documents.

Different definitions of *accuracy* are provided by the data quality literature. In [25], accuracy is defined as "the extent to which data are correct, reliable and certified". In [21], accuracy is associated with data values and is defined as a measure of the proximity of a data value *v* to some other value *v'* that is considered correct. It is simple to determine the level of accuracy of a data value if it represents a characteristic of a real-world entity, which can be used as a source of correct information. If data values are the output of a complex transformation process, accuracy may be difficult to evaluate, as it involves the dynamic assessment of a process, as opposed to the static comparison of values.

However, a measure of accuracy can be associated with data sources. It can be defined as the ratio between the number of correct values and the total number of values available from a given source [21]. This definition supports a mathematical measure of accuracy that does not directly apply to web pages, as

they contain not only elementary data values, but also complex information objects. An indirect evaluation procedure may compare each information object in a page with a corresponding correct copy, in order to identify value mismatches and consequent inaccuracies. If a section of a web page is a result of queries, the evaluation process is more complex, since data may have been extracted with information retrieval techniques and may not come from a single source. In these cases, accuracy can be measured as the distance between the result of information retrieval techniques and a theoretically exact query answer obtained from ad hoc information systems with semi-structured data. In this case, accuracy represents a measure of *precision*, which is the probability that results are relevant for users [26].

The definition of data *completeness* is consistent across research contributions. In [21] completeness is associated with data values and is defined as the degree to which a specific database includes all the values corresponding to a complete representation of a given set of real world events as database entities. According to this definition, it is possible to obtain an objective measure of the completeness of a data source by adding up significant data values.

In web systems, completeness can be measured as the degree to which a web page includes all relevant information. The evaluation procedure is different for documents and elementary data items. If a section contains a document, the corresponding value of completeness can be obtained from the document's metadata that are data certifying the quality of the document. If a section contains a list of elementary data items, completeness can be measured by comparing the list with a data source that is certified as complete.

Currency is not provided a standard definition in the literature. Currency is usually defined as a time interval. In [7] currency is defined as a time interval resulting from merging two time intervals: the first ranges from the time when data are stored in the database to the time when data are used, while the second, referred to as the *age* of data, indicates how old data are when they are stored in the database. Instead, in [9] currency is defined as the time interval ranging from the time when data are updated to the time when data are used (Figure 1). For our purposes, it is important to consider the last update of data in order to evaluate the actual currency of web pages.

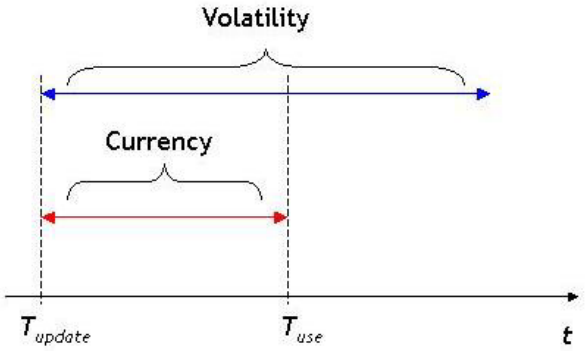

Figure 1: Relationship between volatility and currency.

An important property of data is *volatility* that is the average time length for which data remain valid [7]. Volatility is considered a static property that is independent of architectural design choices. For example, the quotation of a stock remains valid for only a few seconds irrespective of architectural choices. When a page has to be replaced with a new one, it is necessary to evaluate the validity of the old page and, if it is valid, it could still be useful for a subset of users. Furthermore, currency should always be smaller than volatility in order for users to access timely (or up-to-date) data (Figure 1), that is:

$$Volatility - Currency > 0 \qquad (1)$$

In the data quality literature, the focus is on evaluating the quality of current data. Even when a process-based approach is taken, as in TDQM [24], when the production process of information is linked to data quality attributes, only current values are considered.

Data are associated with quality metadata, specifying their accuracy, currency, completeness, consistency and volatility. We suppose that metadata also include an evaluation of authenticity and credibility of data, which are considered fundamental for preservation [15]. Authenticity and credibility are used to guarantee that a page has not been corrupted, which constituted a precondition to preservation.
A web page has a structure specifying the information objects composing the page and their position (Figure 2). Objects can be of different types. For example, HTML is the standard language to create web information objects, but popular document formats such as DOC, PDF and PS can also represent web information objects. Web information objects can also be specified as queries on a specific database. Finally, multimedia files, such as images, sounds and videos can be information objects of a web page [1].

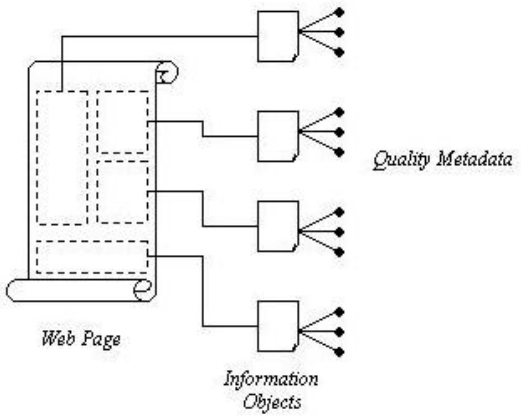

Figure 2: Structure of web pages and related metadata.

Given the diverse nature of information objects, quality metadata cannot be associated with web pages, but must be specifically designed for different types of information objects. A page's quality metadata can be derived by aggregating corresponding objects' metadata according to the page's structure. In this version of our methodology, we only consider documents and data extracted through queries as information objects. The Quality Factory described in Section 5 calculates metadata for data extracted through queries. Documents are supposed to be associated with metadata when they are created.

4. A MODEL FOR PRESERVING THE WEB

This section discusses a preservation process based on the data quality dimensions discussed in the previous section. The process is based on the *information management life cycle* defined by Gail M. Hodge [15] within the digital archiving context. The information management life cycle is composed by the following subsequent steps, each of which is associated with corresponding critical issues:

- *Creation* – The creation phase produces information. Who produces information is considered responsible for its quality. Therefore, quality metadata are introduced in this phase, to be used in subsequent phases. In particular, who creates information should also estimate the long-term value of the information, expressed in terms of volatility. However, in many cases, metadata are created by measurement or monitoring instruments and automatically associated with data. This instrument-generated metadata can be supplemented by information provided by the original creator of the information.
- *Acquisition* – Acquisition incorporates a new information object into the digital archive. Selection is considered a step of the acquisition phase, as it is necessary to limit acquisition to information objects of high cultural and research value. In order to perform selection, it is necessary to make decisions on the depth with which a page's links have to be explored and evaluated for preservation.

- *Cataloguing* – After acquisition, it is necessary to catalog information objects. In this phase, there are two main critical issues. The former is related to the representation of metadata and their level of detail. The latter is the need for persistent identification that is the choice of a unique identifier for the new information object.
- *Storage* – Storage inserts new information objects into the digital archive. A fundamental issue at this stage is the choice of the storage media. For example, the literature reports case studies about irretrievable data on 5¼" floppy disks, which cannot be read by the modern PCs. Retrieving those data involves a cost, which, conceptually, represents a loss from wrong storage decisions.
- *Preservation* – Preservation is concerned with management activities that preserve the content as well as the look and feel of information objects. In particular, the format in which data have to be preserved is chosen at this stage, which may result in a need for transforming data before preservation in order to have a uniform document format in the digital library.
- *Access* – Access ensures continuous accessibility of the information objects stored in the digital archive.

From our perspective, the main problem with this approach is that it may not be feasible to preserve all online information. As we have discussed in Section 2, we assume that data owners have the responsibility for preservation. These decentralized responsibilities share the preservation burden and, if they follow a common preservation procedure, they can preserve large data sets that would exceed the capacity of a centralized preservation unit. Under this assumption, the preservation process starts with data creation. Each time a new page has to be published, it is necessary to execute the procedure shown in Figure 3, named s*tatic preservation model*. In literature [5] preservation strategies are classified into three categories:

- Preservation of bits: the goal is the preservation of the exact bit sequence of original data.
- Preservation of content: the goal is the preservation of the content, such as the words of a text or the appearance of an image, without storing information on how this content is presented to users.
- Preservation of experience: the goal is the preservation of the dynamic and interactive nature of web information.

In this paper, the goal is the preservation of experience, therefore a page is preserved with all its information objects and structure. A web page is considered to be updated even if only one information object is modified, as this modification can affect the structure of the page. A modified page is considered a new object to be preserved (Figure 3).

At creation time, data are associated with metadata, describing their quality. Quality is expressed in terms of accuracy, completeness, consistency, currency, and volatility as described in the previous section. Metadata also include general properties of the document, such as the author and the document type. Metadata are used to decide whether to preserve a document after each update and, if it is preserved, whether to attribute to the updated version of the document a new URL (see Figure 3).

After creation or update, a page is published and, at the same time, it is acquired as an input by the preservation cycle. Before the acquisition phase is executed, the quality of the web page is evaluated, along the accuracy, completeness, consistency, currency, and volatility dimensions. The user specifies minimum acceptable values for all quality dimensions and, if new data satisfy quality requirements, they will be physically or virtually incorporated into the archive. After acquisition, the web page is catalogued. It is associated with a new inventory number, missing mandatory metadata are added and then the page is stored in an appropriate physical support. If evaluation results do not meet quality requirements, data are returned to their owner with a warning and are not catalogued until their quality is increased or quality thresholds are decreased.

As regards the preservation stage, it is important that web pages include three different types of documents: HTML, PDF/RTF or multimedia. Considering the case study described in this paper, we propose to preserve PDF/RTF and multimedia documents either as is or compressed. As regards web pages, we suggest to translate HTML pages in XML pages. XML has many advantages, such as

flexibility in metadata management and the ability to manage an archive of documents as a database, by using query languages to extract a specific document.
In the publishing stage, the author of new data is sent the URL at which data are published. When a new web page is published by replacing an old web page, the residual importance of the old page has to be evaluated. Importance is evaluated against the volatility dimension. If volatility is low, the risk of information loss increases, since data are often updated; on the contrary, if volatility is high, the information is static and it can be supposed that information losses due to updates are less likely. If evaluation results indicate that old data are still valid, data are not deleted, but they are associated with a new URL and linked to the same web page as updated data.

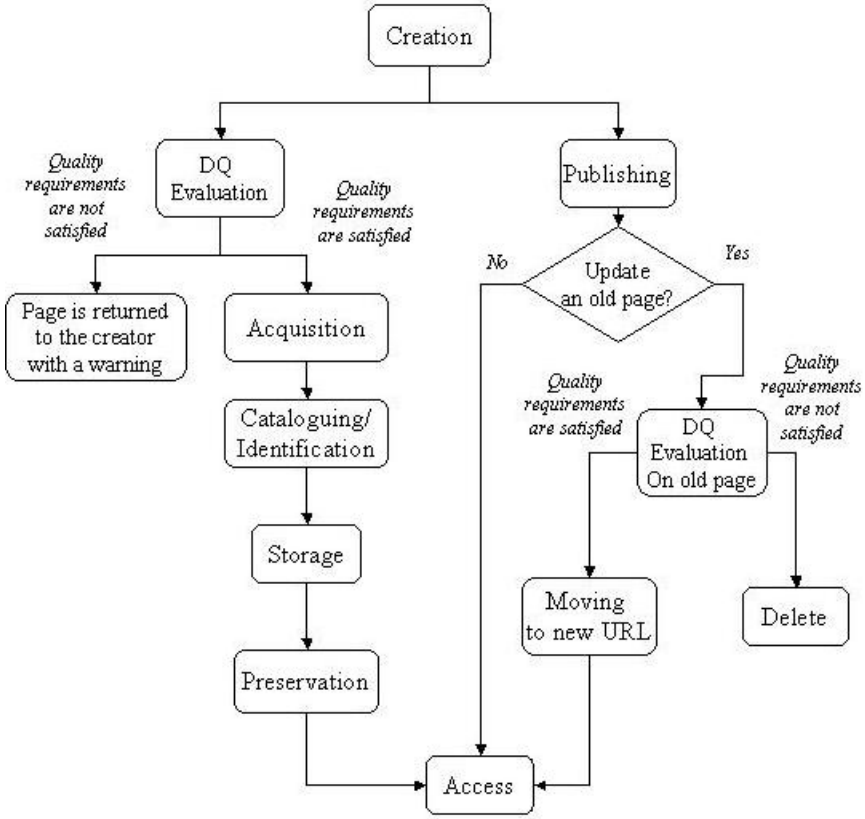

Figure 3: Static preservation model.

In order to increase the effectiveness of preservation and the availability of useful information, a dynamic procedure should be implemented to monitor web pages periodically. This dynamic preservation model controls the validity of web data by evaluating the volatility dimension. If data have expired, they are deleted and the structure of the web site has to be updated. If data are still valid, their URL is confirmed (see Figure 4).

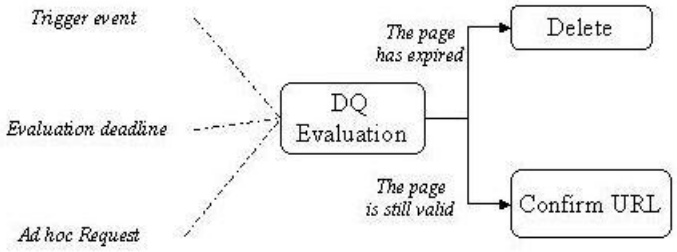

Figure 4: Dynamic preservation model.

Data quality evaluations involved in the models described above are performed by a software module that is discussed in next section.

5. THE QUALITY FACTORY

The methodology described in the previous section assumes that a central software module, called Quality Factory (QF), is responsible for the evaluation of information quality. Figure 5 shows the main modules of the Quality Factory and their interactions. The Quality Factory is composed by four modules: *Quality Analyzer*, *Quality Assessment*, *Monitoring and Data Quality Certificate*.

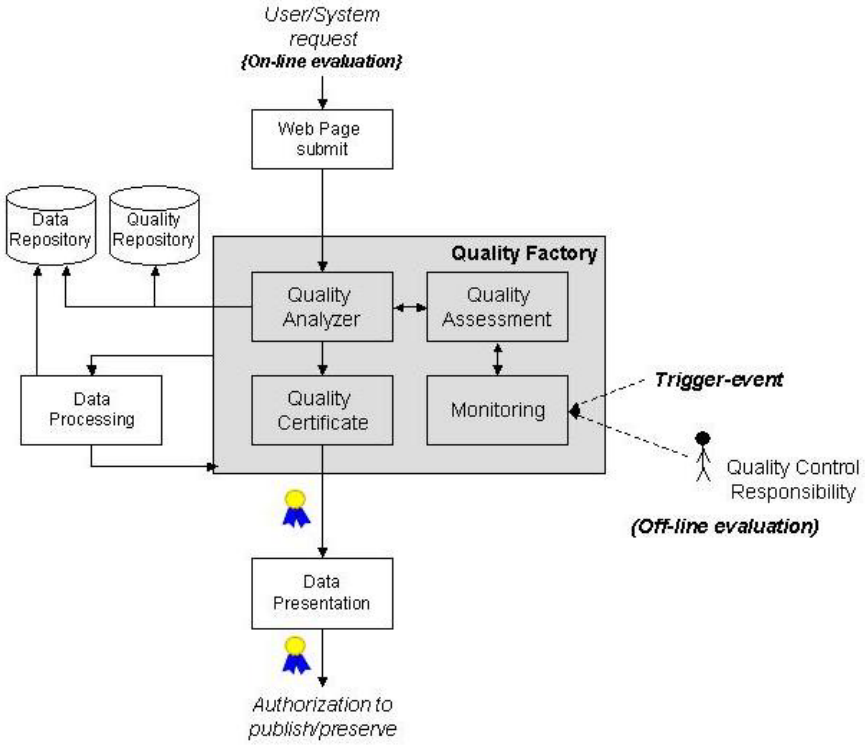

Figure 5: The Quality Factory

Evaluation can be performed online or offline. If evaluation is performed online, web pages are submitted to the QF, which evaluates quality in real time. Submissions correspond to both a *publish request* and a

preservation request. With publish requests, users can specify the minimum value that is considered acceptable for the quality dimensions to be evaluated.
Quality Analyzer analyses the page and its content and, if necessary, extracts a *reference web page* from the Data Repository. The reference web page is used to obtain a comparative evaluation of the quality of the submitted web page. Information about data quality dimensions is stored in the *Quality Repository* and is extracted from the quality analyzer when needed. The measure of different quality dimensions is executed by the *Quality Assessment* module, which uses a set of measurement tools. The result of the evaluation process, i.e. the quality metadata, is returned together with the submitted web page.
If a web page does not satisfy quality requirements, the *Quality Assessment* module sends an alert message to the *Monitoring* module. The message contains the web page that does not satisfy quality requirements and corresponding critical quality values. If quality values are satisfactory, the *Quality Assessment* module notifies the good quality of the information to the *Quality Analyzer* and a quality certificate is associated with web page by the *Quality Certificate* module. The certificate has two different meanings as it represents a response to two different requests. As a response to publish requests, the certificate represents a license to publish the web page. As a response to preservation requests, the certificate measures the benefits from preserving the web page.
Finally, quality evaluation data are sent to the *Data Processing* module that cooperates with other software applications that are responsible for the final response to the user. In particular, the *Data Presentation* module is responsible for sending the response to the user in a specific format through a *Quality Service Interface*.

The offline evaluation differs from the online evaluation since it is not triggered by the submission of a web page. The QF periodically calculates the quality metadata associated with the web pages of an organization's site, to support preservation operations. In this way, the system automatically identifies all web pages that should be preserved. The offline evaluation can also be event-based. The *Monitoring module* can store the rules to identify the events that activate the offline evaluation. When an event occurs, the *Monitoring module* sends a message to the *Quality Assessment module* to start the evaluation process. Quality metadata are stored in a specific database to be retrieved by subsequent preservation procedures.

6. CASE STUDY

The model discussed in the previous sections has been tested on Politecnico di Milano's web site. Only the main domain (www.polimi.it) has been considered for testing, while secondary domains that are independently managed, such as departments' web sites, have not been analyzed.

Politecnico's web site has a number of features typical of scientific and cultural sites, which differentiate it from commercial sites. As a general observation, users cannot be considered customers and the site offers advanced services without direct economic returns. These features encourage preservation, as increasing the quality of the site is a goal even if it does not deliver immediate returns. Nonetheless, the feasibility of the preservation model has been verified by performing a risk analysis, an assessment of the investment and an evaluation of the impact on the site's management procedures.
Risk analysis has been conducted by analyzing the causes for information losses and their consequences. Technical failures and updates represent the main source of information losses. Technical failures could be avoided by means of back up and fault tolerance techniques. The latter would require a change in update procedures, which should not overwrite previous web pages and documents. Replacing an old page with a new one is simple, as it avoids changes in the site structure. Preserving all versions of documents would also increase the memory size of the site and, thus, raise costs. Resulting information losses have been classified as follows:

- A new document is created: pages have a finite validity and periodically new versions are created to replace them. In many cases, deleting old versions of pages could represent an information loss. For example, the teaching program of the school is periodically modified and, in general, the version available online is the one valid for the ongoing academic year. Over the past nine years, Politecnico di Milano has changed the official teaching program four times. The site publishes only the last version, although elder students must follow the previous version of the program. As a consequence, they have to rely on paper versions of the old teaching program, which is available from the students office.
- New information is added to a page, by either extending or replacing previous information. Replacing previous information can generate losses of useful data.
- An old but still useful document is deleted. The site has been monitored and many pages have been deleted without creating corresponding new pages.

It has been verified that there is no reference model to establish which information has to be preserved and when web pages can be deleted from the site. Designing and implementing a preservation model would involve costs that are related to the following new resources:

- Human resources: at least one full-time expert responsible for data quality and a temporary implementation team to design the software are required for preservation. However, in a more general context, preserving web information needs not only IT specialists, but also staff from the library with archiving professional skills.
- Physical resources: the hardware and software infrastructure needs to be upgraded to support preservation.

On the other hand, the preservation process would simplify secretarial work, by reducing the time needed to find old documents requested by students, which have been deleted from the web site. It has been estimated that the number of students who could require documents that have been removed from the site evaluates to 10.500 that is about 40% of the university population. This datum is provided by Politecnico of Milano's Students Office and refers to the academic year 2002/2003.

Interviews have been conducted to evaluate the impact of preservation on site management procedures. A questionnaire has been submitted to a sample of Politecnico's employees with different roles in the preservation cycle. The questionnaire is composed by 25 questions that evaluate the perceived need for a preservation procedure. An aggregate index of preservation need has been calculated based on answers as the mean value of interviewees' judgments. The index evaluates to 4,12 out of 5 and indicates a high need for a preservation procedure. Furthermore, results indicate that the model is perceived to have a limited impact on site management activities, as most of the effort is related to preparing, as opposed to publishing documents.

The application of the model requires an initial phase to analyze the information of the web site in order to operate a selection and make decisions on preservation criteria and quality thresholds. Then, the full-time data quality responsibility has to perform monitoring activities and trigger event management actions. As described in Section 4, the selection of documents that have to be preserved is executed on the basis on data quality dimensions, such as volatility. In our experimentation, documents available online are listed and classified according to the probability with which they could generate information losses. Three classes of risk are identified:

- Not risky: information that has a very high volatility (five years and above) is not changed frequently and the probability to lose information as a consequence of updates is very low.
- Potentially risky: information that has a medium volatility (between one and five years) and could be subject to updates and consequent information losses.
- Risky: information that has a low volatility (below one year) is frequently updated and is very likely to generate information losses.

Documents available online are about 1,056 and about 19,9% belong to the "not risky" category, 16% belong to the "potentially risky" category, 48,1% belong to the "risky" category, while the remaining 16%

represent links to external information. These percentages indicate that the amount of information that could be lost as a consequence of updates is significant.

7. CONCLUSIONS

The paper shows how a data-quality approach can be applied to the preservation of web information as a means to select a subset of high-quality of web pages to be preserved. Quality thresholds that determine the selection of information for preservation are user defined and, thus, context dependent. A fundamental benefit of the contextual nature of data quality techniques is that, although the evaluation of quality is performed automatically, the preservation process remains under the control of experts with archival skills. This reflects a fundamental principle of traditional libraries of paper documents, where librarians base their judgments both on objective measures of the relevance of information, such as the number of borrowers, and on their own contextual experience with library preservation procedures.
Future work will consider a more accurate testing of the preservation approach. This will be achieved by implementing the preservation procedure in a real case and measuring the outcomes over time, in terms of both users' satisfaction and overall quality of the site over time.

ACKNOWLEDGEMENTS

This work has been partially supported by the Italian MIUR-MURST COFIN DaQuinCis project and by the Italian FIRB Project MAIS. Particular thanks are expressed to Luca Ciccarelli for his assistance in model definition and data collection activities.

REFERENCES

[1] Abiteboul, S. Issues in Monitoring Web Data. *Proceedings of the 13th International Conference on database and Expert Systems Applications,* 2002, France. Lecture Notes in Computer Science, 2453. Berlin: Springer, 1-8.

[2] Abiteboul, S., Cluet, S., Ferran, G., Rousset, M.C. The Xyleme Project. *Computer Networks,* 39 (3), pp. 225-238.

[3] Abiteboul, S., Cobéna, G., Masanès, J., Sedrati, G. A first Experience in archiving the French Web. *Proceedings of Research and advanced technology for digital libraries: 6th European Conference, ECDL 2002,* Italy. Lecture Notes in Computer Science, 2458. Berlin: Springer, 1-15.

[4] Alexa Internet. Available online at http://www.alexa.com/

[5] Arms, W.Y., Adkins, R., Ammen, C., Hayes, A. Collecting and Preserving the web : The Minerva Prototype. *RLG DigiNews*, vol. 5 No. 2, April 2001. Available online at http://www.rlg.org/preserv/diginews/diginews5-2.html

[6] Ballou, D. P., Pazer, H.L. Modelling Data and Process Quality in Multi-input, Multi-output Information Systems. *Management Science*, vol. 31, No. 2, February 1985.

[7] Ballou, D. P., Wang, R., Pazer, H.L., Tayi, G.K. Modelling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, vol. 44, No. 4, April 1998.

[8] Ballou, D. P., Pazer, H.L. Designing Data Information Systems to Optimize the Accuracy-timeliness Tradeoff. *Information Systems Research*, 1995.

[9] Bovee, M., Srivastava, R.P., Mak, B. A Conceptual Framework and Belief- Function Approach to Assessing Overall Information Quality. *Proceedings of the Sixth International Conference on Information Quality*, 2001.

[10] Bury, S. *Domain.uk: interim report.* London British Library, March 2002 (Internal Report)

[11] Cappiello, C., Francalanci, C., Pernici, B. A Model of Data Currency in Multi-channel Financial Architectures. *Proceedings of the Seventh International Conference on Information Quality (ICIQ '02)*, Boston, MA, USA, 2002.

[12] Cappiello, C., Francalanci, C., Pernici, B., Plebani, P., Scannapieco, M. Data Quality Assurance in Cooperative Information Systems: a Multi-dimension Quality Certificate, *Proceedings of the ICDT 2003 International Workshop "Data Quality in Cooperative Information Systems" (DQCIS 2003)*, Siena, Italy, 2003

[13] Day, M. *Collecting and preserving the world wide web*. A feasibility study undertaken for the JISC and Wellcome Trust, UKOLN, University of Bath, Version 1.0, February 2003. Available online at http://www.jisc.ac.uk/uploaded_documents/ archiving_feasibility.pdf

[14] ERPANET Project. Available online at http://www.erpanet.org

[15] Hodge, G.M. Best practises for digital archiving, *D-Lib*, January 2000. Available online at http://www. Dlin.org/dli/January00/01hodge.html

[16] InterPARES Project. Available online at http://www.interpares.org

[17] National Library of Australia. Available online at http://www.nla.gov.au/

[18] Orr, K. Data Quality and Systems Theory. *Communications of the ACM*, vol.41, no.2, February 1998

[19] Orr, K. *Data Warehousing Technology*. The Ken Orr Institute 2000.

[20] Redman, T.C. The Impact of Poor Data Quality on the Typical Enterprise. *Communications of the ACM*, vol. 41, no. 2, February 1998.

[21] Redman, T.C. *Data Quality for the Information Age.* Artech House, 1996.

[22] Rothenberg, J. Ensuring the longevity of digital information. *Scientific American* 272 (1):24-29, January 1995.

[23] Wand, Y., Wang, R.Y. Anchoring Data Quality Dimensions in Ontological Foundations. *Communication of the ACM*, vol. 39, no. 11, 1996

[24] Wang, R.Y. A Product Perspective on Total Data Quality Management. *Communications of the ACM*, vol. 41, no.2, February 1998.

[25] Wang, R. Y., Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, Spring 1996, Vol.12, No.4, pp.5-34.

[26] Weikum, G. *Towards Guaranteed Quality and Dependability of Information Services*. In Invited Talk at the 8th German Database Conference (Datenbanksysteme in B uro, Technik und Wissenschaft), 1999.