A Framework to Evaluate Information Quality in Public Administration Website

Filippo Geraci
Istituto di Informatica e Telematica, CNR
Via G. Moruzzi, 1 - 56124 Pisa, Italy
filippo.geraci@iit.cnr.it

Maurizio Martinelli
Istituto di Informatica e Telematica, CNR
Via G. Moruzzi, 1 - 56124 Pisa, Italy
maurizio.martinelli@iit.cnr.it

Marco Pellegrini
Istituto di Informatica e Telematica, CNR
Via G. Moruzzi, 1 - 56124 Pisa, Italy
marco.pellegrini@iit.cnr.it

Michela Serrecchia
Istituto di Informatica e Telematica, CNR
Via G. Moruzzi, 1 - 56124 Pisa, Italy
michela.serrecchia@iit.cnr.it

Abstract

This paper presents a framework aimed at assessing the capacity of Public Administration bodies (PA) to offer a good quality of information and service on their web portals. Our framework is based on the extraction of ".it" domain names registered by Italian public institutions and the subsequent analysis of their relative websites. The analysis foresees an automatic gathering of the web pages of PA portals by means of web crawling and an assessment of the quality of their online information services. This assessment is carried out by verifying their compliance with current legislation on the basis of the criteria established in government guidelines[1]. This approach provides an ongoing monitoring process of the PA websites that can contribute to the improvement of their overall quality. Moreover, our approach can also hopefully be of benefit to local governments in other countries.

Keywords: E-government, websites, public administration, crawling

[1] http://www.funzionepubblica.gov.it/media/835828/linee_guida_siti_web_delle_pa_2011.pdf

Introduction

Within a context of global economic crisis, such as the present one, continuous innovation is one of the most effective levers to use in supporting development. The introduction of new technologies, the re-engineering and simplification of processes must be widespread at all levels, beginning with local government bodies. It is within this context that the European Union has set up a series of initiatives among which there is the European Digital Agenda. Its aim is to build a new "digital administration" able to respond more rapidly to the needs of citizens and enterprises also by means of the provision of e-government services through the web portals of local PA bodies.

This paper proposes a methodology to measure the ability of public administrations to create a high quality offer of information and services by means of the analysis of the web portals. This work has been carried out in the framework of cooperation with the "Department for Digitalization of the Public Administration and the Technological Innovation" of the Presidency of the Italian Council of Ministers. Our choice to use institutional websites as an endogenous metric for the assessment of the local government bodies has many advantages. The most important advantage of endogenous metrics is that they are much less subject to bias when compared with traditional exogenous metrics such as market researches (Diez-Picazo, 1999) because they are based on automatic data gathering and retrieval. In addition, they allow good geographical characterization of a phenomenon of interest since they are based on data that allows the differentiation of users on a national, regional and provincial level (Martinelli et al., 2006).

Our framework is based on three main steps: 1) extraction of the list of .it PA domains to be analyzed, 2) automatic gathering of web pages from their web portals by means of a web crawler and 3) analysis of the downloaded data with the aim of producing a quantitative evaluation of the PAs websites content.

The selection from the list of domain names to analyze depends on its specific context. Some ad hoc solutions can be used, while other problems are recurrent. However, this step is fundamental for an effective conclusion. The main issue to be faced is that of guaranteeing the complete collection of the data, making sure that the sample being examined represents the real situation. In other words, we must be sure that the list of domain names is comprehensive (namely it contains the domains of all the PA bodies of interest in the analysis). This is to limit the problem of underestimation of the phenomenon being studied. In fact, the use domain names as a metric, besides the advantages mentioned above, such as the automatic determination of registrant characteristics (individuals, companies, public administrations, non-profit institutions) and geographical characterization (national, macro-area, regional and provincial) of the bodies examined has also some drawback. These mainly regard under- and over-estimation. Underestimation occurs when the institutions register their domain names within a gTLD (e.g. .com, .biz, etc.) or another ccTLD (e.g. .eu). On the other hand, we can have overestimation when more than one domain name is registered by the same administrative body. In this analysis, besides paying attention not to leave out some public bodies, we must also consider that for each institution, all the relevant domains must be included to avoid underestimation of the number of compliant websites.

The term web crawling refers to the activity of discovery and automatic downloading of large portions of the Web. Crawling software is made up of a set of interconnected modules. Among the most important modules there are: the parser, which analyses the downloaded documents and extracts hypertext links (here referred to

as URLs) to other pages; a document database containing the URLs identified but not yet downloaded; a database for the management of the information regarding the crawling progress for each website (downloaded pages, failures, web analytics, interconnection with other pages (web graphs); an index of downloaded pages; a module for page downloading.

The first step to evaluate the quality of information of a PA is that of verifying the presence of a minimum set of contents inside its institutional portal and classifying each content item on the basis of how it is easy to identify and reach it. The identification and accessibility of the content depends on various factors. These include the presence of a direct link on the home page or in a page directly reachable from the home page, the use of unambiguous or misleading anchors, etc. The contents of the websites have been classified in three possible categories: 1) compliant, when the content is perfectly identifiable and reachable, 2) not compliant, when the content is on the website, but is not directly accessible, 3) missing, when the content has not been detected within the portal.

Our framework has been tested by evaluating Italian local government institutions by extracting the list of domains registered in the ccTLD ".it" Registry, managed by the Institute of Informatics and Telematics of the CNR (National Research Council) of Pisa, Italy. The access to this database allowed us to obtain the complete list of domain names and to avoid, therefore, any loss of information and consequent distortion of results. As far as the analysis is concerned, we assessed the compliance of the local government websites with the guidelines issued by the directive of the Ministry of Public Administration and Innovation within the e-Government Plan. The aim of the guidelines is to provide indications regarding the general criteria and operating tools for the rationalization of online contents for the Italian public administrations. In particular, the guidelines pay special attention to the definition of minimum contents set that must be included within the institutional websites of local government bodies.

The paper is organized as follows: in section 2 we examine the literature related to the crawling problem and the evaluation of e-government web contents; in section 3 we give more details about the architecture of our framework, in section 4 we show our experimental results which importance is discussed in section 5. Section 6 draws conclusions.

Related Work

Since crawlers are the most important tools for mining data from the web, a lot of different solutions are described in the literature. A complete description of much well known crawling architecture can be found in the Ph.D. Thesis of Carlos Castillo (Castillo, 2004). General issues related to web crawling are described also in surveys such as that of (Olston and Najork, 2010). Web crawling ethics is discussed in (Thelwall and Stuart, 2006). Strategies for re-crawling the web so as to maintain high freshness are discussed in (Edwards, 2001). Strategies for URL selection for attaining high quality pages are discussed in (Boldi, 2004). The queues of detected but as yet unvisited URLs is one of the most important and dynamically changing data structures in a crawler and many optimizations have been proposed in (Marin, 2008).

Many models have been proposed in the literature aimed at improving the communicative interchange between citizen and government agencies. In this respect, (Wang and Bretschneider, 2005) presented a model that not only evaluates web-based e-government services, but also helps government agencies to understand why their websites succeed or fail to help citizens in finding requested information. In (Fong and Meng, 2009), the authors state that an e-Government portal represents not only a public image of a sovereign region, but also a reliable service platform for many users from local citizens and beyond. This

makes the requirement for robustness of an e-Government portal relatively stringent. The authors propose a web-based performance monitoring system (WMS) for checking the health status of the service portals in real-time. Specifically, the authors discuss the context of applications of WMS to e-Government performance models. (Alshehri et al., 2012) discussed an investigation of the effect of the Website Quality (WQ) factor on the acceptance of using e-government services (G2C) in the Kingdom of Saudi Arabia (KSA) by adopting the Unified Theory of Acceptance and Use of Technology (UTAUT) Model. Survey Data collected from 400 respondents were examined using the structural equation modelling (SEM) technique and utilising AMOS tools. This study found that the factors that significantly influenced the Use Behaviour of e-government services in KSA (USE) include Performance Expectancy (PE), Effort expectancy (EE), Facilitating Conditions (FC) and Website Quality (WQ). (Moon, 2002) examined the current state of municipal e-government implementation and assesses its perceptual effectiveness. This study also explored two institutional factors (size and type of government) that contribute to the adoption of e-government among municipalities. Overall, this study concluded that e-government had been adopted by many municipal governments, but it was still at an early stage and had not obtained many of expected outcomes (cost savings, downsizing, etc.) that the rhetoric of e-government had promised. (Sousa and Lopez, 2002) used a municipal e-government services framework to evaluate the development of municipal electronic government (e-Government) in Peruvian cities by assessing the web sites and services provided by each city. The findings showed that the degree of development of e-Government in Peru is incipient, with the need to improve not only the municipal websites but also the municipal e-Government strategy in each city within a common framework. Finally, the authors compared the results obtained in Peru with other Ibero-American countries and, in general terms, this comparison reveals a poor level of e-Services development in Peru in comparison with its neighbors who are more developed. (Torres et al., 2005) studied the development of e-government initiatives at the regional and local level in the EU through the opinion of those agents directly involved in the projects. The authors wrote a questionnaire that was sent to the regions and the largest cities of EU countries, in order to find out their degree of involvement in e-government initiatives. Responses were received from 47 regional and local governments. The survey findings showed that e-government initiatives are still predominantly non-interactive and non-deliberative. (Scavo, 2003) analyzed 145 municipal and county government websites. These data are used to examine how local governments were using the Web and to examine the evolution of Web usage over the four years between the first survey conducted in 1999 and the second survey conducted in 2002. The author concluded that local governments have made progress in incorporating many of the features of the Web but that they have a long way to go in realizing its full promise.

Our Approach

The objectives of the framework proposed in this paper can be summarized in: 1) creation of an experimental methodology of evaluation of the quality of PA websites and measurement of the level of use and efficiency of services; 2) provision of a tool for the analysis of specific phenomena relative to the process of digitalization of Public Administration (e.g. respect of the minimum requisites of public websites, as required by current legislation and by the Guidelines issued by the Ministry of Public Administration and Innovation) and of their geographical characterization; 3) provision of a tool for the ongoing observation of the PA websites; 4) give support to any initiatives involved in the continuous improvement in the quality of PA websites. In order to reach these objectives three main steps were taken: extraction of the

entire domain names to be analyzed, automatic gathering of web pages from the portals by means of crawling and analysis of these data to produce a quantitative assessment of their information content. The choice to carry out the evaluation of the contents offline has many advantages. The most important one is that, storing the downloaded data, we can perform new analyses every time a new need arises. Moreover, this allows comparison between the past and present status of any single website.

Extraction of Domain Names from the ccTLD ".it" Database

Domain names registered under the ".it" ccTLD are extracted from the database of registrations managed by the Institute of Informatics and Telematics (IIT) of CNR, Pisa, using automatic and semiautomatic procedures. The .it Registry subdivides the domains into seven categories (individuals, firms, professionals, non-profit institutions, public bodies, other institutions and foreign entities). Particular attention was paid to the registration of domain names by public bodies. From the database of domains labeled as public bodies, the local institutions were extracted, in particular the bodies of "Municipality", "Province" and "Region", which reflect the basic public administration structure in Italy. Furthermore, these institutions were classified according to their geographical location (at the level of macro area, regional and province). In this phase also semi-automatic checks were carried out in order to avoid distortions of results. During the first stage of skimming, some errors of registration classifications were detected. These were both errors in classification of the institutions (e.g. public bodies classified in the database as non-profit organizations, or some private companies wrongly categorized as public bodies) and also in the classification of bodies on a geographical basis (some bodies were associated with the wrong region or province). These errors were eliminated so as to obtain qualitatively reliable data.

Crawling of the Websites of the Public Administration

Crawlers are systems whose purpose is to build a local image of the structure and content of (a portion of) the Web. The crawling task is a preliminary step for the activity of analyzing the Web. The main advantage of keeping a local image of the Web is that it is always possible to make further investigations on the same data if necessary, without introducing computational costs (it involves only the additional cost of data storage).

The characteristics of the Web require a careful design of the architecture of web crawlers. Even for a small fraction of the Web such as that published by public administrations, the overall size of the data and the number of documents to be downloaded requires crawlers to have a parallel architecture, to implement ad hoc solutions for the organization of data in primary and secondary memory and to carry out specific crawling policies. Moreover, crawlers must be able to address and recover network and software failures, so preserving data integrity and preventing data loss.

In designing our crawler (Felicioli et al., 2011) we took into account all these issues by designing a modular architecture in which data management, crawling strategy, network management and the other tasks are separate. The advantage of this architectural choice is twofold: it ensures a logical separation between different tasks, and it makes the crawler expandable and configurable. Our crawler has a parallel architecture that can run on a configurable number of standard workstations, each of which runs an instance of the crawler. Each instance is responsible for a subset of hosts and maintains locally all the information about them. Once a new URL is found it is sent to the appropriate instance of the crawler that will manage it. The assignment of hosts to a workstation is performed through a hash function. This allows us to avoid centralized operations.

In order to obtain a high crawling speed it is essential to minimize disk access. To achieve this goal we designed a model in which there are two types of data: blocks and logs. Blocks contain frequently accessed data, have a fixed length and are organized in size-classes. This type of data is used to maintain general information about each host. In Figure 1 we show the internal organization of a block. Each block contains a pre-allocated number of slots (depending on the block's size-class) to maintain fixed size data of each page of the host (a unique identifier, the crawling status, etc. and a pre-allocated space used to memorize the list of the URLs of the pages. If during crawling the number of discovered pages exceeds the number of available slots, the block is copied into another of higher size-class. The same happens if the size of the list of URLs exceeds the available space. When downloading a certain website the related block is maintained in RAM. This allows a reduced computational cost when the block must be copied in a higher size-class. Once the download is complete, the block is permanently stored on disk. Logs contain data are no longer required during crawling (i.e. HTML pages). Logs are buffered in the main memory and periodically stored on disk.

Since downloading all the pages could be impractical for huge or dynamic websites, crawlers spend a limited pre-defined amount of resources and time collecting a sub sample of each website. In these cases the crawling strategy has a deep impact on the significance of the website sample. We implemented an iterative crawling strategy so that for each host we maintain a statistically meaningful sample of a website. To achieve this goal we use a crawling strategy which is a hybrid of DFS and BFS.

Another important issue we addressed is the tolerance to network and software failures. To achieve this result we preferred to do not download an entire host in a single pass, but we used an iterative approach so that we download a limited number of pages per host at each step and store the partial result (its block and its log) on disk. At each iteration on the same host, its block is reloaded from disk. The advantage of this policy is that the failure of a crawling thread causes a limited data loss (only the pages still in RAM) and consequently, a limited time overhead to re-download the missing pages. Moreover, we designed data structures in secondary memory to maintain consistency. To achieve this goal, we used techniques similar to those used for journaling file systems. In a nutshell, when we write a block on disk, we do not overwrite the previous block, but we write it in a free space. Then we update the journal modifying the pointer to the block and marking the previous block as free space.

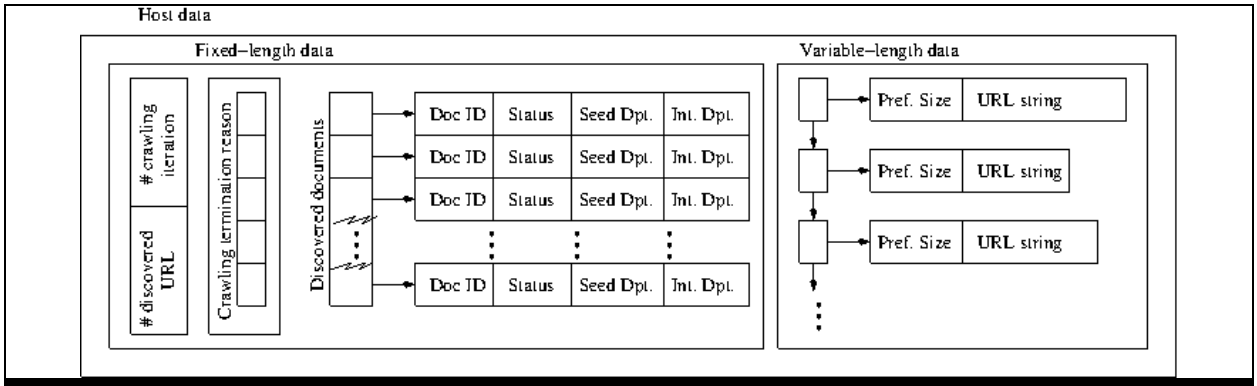


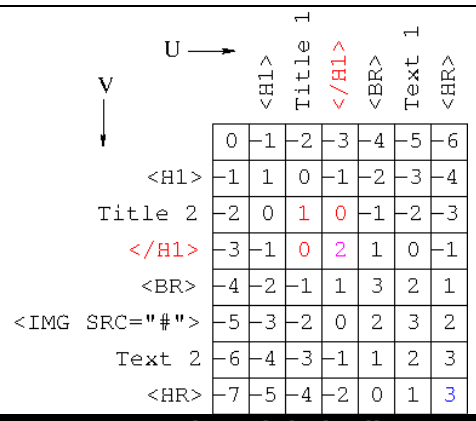
Figure 1 - Internal organization of the data structure used to memorize host data

Identification of the Website Associated to a Domain

Running a crawler over a set of domains can return an unexpected result. In fact, the website related to a certain domain does not necessarily correspond to the domain itself. More often web servers are configured to require the prefix “www” or some other non-standard prefix. In practice, two or more admissible website base URLs can return a valid HTML page and these pages can be different. For example, we observed that the domain with the “www.” prefix is preferred for the institutional portal and the domain without prefix is often used for the webmail internal portal. This behavior must be addressed in order to analyze the “correct” website. This means that if the web server returns a valid HTML page with and without the “www” prefix we should download both. A simple equality test on the home page to avoid downloading twice the same web site is too simplistic and fails in most cases. This is due to the fact that succeeding download of the same URL not necessarily returns the same page. For example it suffices that a page contains the counter of the visitors or it contains ads to produce a slightly different HTML. To recognize this situations, we implemented a method (Geraci et al., 2011) based on a well-known dynamic programming global alignment algorithm which compares pairs of home pages in order to discriminate whether different websites associated to the same domain are (near) duplicate websites or not.

In a nutshell the algorithm works as follows: the two HTML pages to compare are split in tokens (a token is a HTML tag or the slice of text between two tags), than we build a matrix in which each position is associated to a pair of tokens (one of a page and one of the other page). Than we populate the matrix assigning a score to each cell from left to right and from top to bottom. The score is a function of the match (mismatch) between the considered tokens and the score of the populated neighbor cells. At the end of this procedure the score in the bottom right cell of the matrix is the global alignment score. We normalize this global score dividing it for the number of tokens of the longer HTML page and thus obtaining a (0,1) ranged value such that it hold 1 if the two pages are identical and 0 if they are completely different. We consider as almost identical two pages if their score exceeds 0.9. Figure 2 shows an example of global alignment.

This choice makes our method robust against local modifications (i.e. counters, advertisements, absolute URLs) that do not affect the global structure of the page. When we identify a domain with two (or more) different websites associated, we have to analyze both before identifying the suitable institutional portal.



		<H1>	Title 1	</H1>	 	Text 1	<HR>
	0	-1	-2	-3	-4	-5	-6
<H1>	-1	1	0	-1	-2	-3	-4
Title 2	-2	0	1	0	-1	-2	-3
</H1>	-3	-1	0	2	1	0	-1
 	-4	-2	-1	1	3	2	1
<IMG SRC="#">	-5	-3	-2	0	2	3	2
Text 2	-6	-4	-3	-1	1	2	3
<HR>	-7	-5	-4	-2	0	1	3

Figure 2 - The matrix to compute the global alignment score. In blue the (not normalized) global score. The score in magenta is a function of the scores in red and of the red tags.

Identification of the Home Page

According to the user experience, the home page of a website is the landing page visualized once one write the base URL of the site in the location bar of her own preferred browser. The identification of the correct home page is an important issue because a certain number of guidelines explicitly require the presence of certain contents in the home page. The identification of the home page from the document root of the website is a quite complicated task. In fact, redirects and aliases must be taken into account. Moreover, redirections can be put in cascade, forming a chain in which the last page is the real website home page.

Aliases are redirections to other websites, which must be crawled and analyses in place of the redirecting website. This rule holds even when the target of the redirect is a website outside the ".it" ccTLD. In fact, it is unfair to state that a certain administration fails to implement guidelines just because its portal is hosted outside the ".it" ccTLD even if it is directly accessible from inside the ".it" domain via the alias.

Identifying redirects is complicated by the fact that web designers very often do not use the standard directive prescribed by HTTP protocol, but generate a valid HTML page which, when interpreted by a web browser, is immediately substituted by another page. We implemented classification software, called tracker, which is able to predict if, once interpreted by a web browser, a web page will produce a redirection and, in this case, extract the target URL. To achieve this goal our software analyzes the internal structure of the page searching for possible typical redirections obtained via: javascript, meta tags or frames.

Quantitative Analysis of Website Information Content

The module for the quantitative assessment of information content has three inputs: 1) a database containing all the web pages of one or more portals downloaded by the crawler; 2) a list of keywords which characterize the content to be detected; 3) a description of the structural characteristics of the page snippet in which the keyword must be found (i.e. an anchor inside the home page pointing to a web page in the same domain).

The database of HTML pages is given as input to software able to extract the textual contents of each page and subdivide them on the basis of the "context" in which they are found. Content extraction is complicated by the fact that it could be contained within images or visualized to the user by means of JavaScript code. The extraction module is based on a parser able to recognize

HTML tags and to evaluate in which cases there is the need to search for content contained within the tag parameters and in which cases it is safe to ignore the tag. In this way it is possible to search content associated to images or inside a JavaScript string. The search algorithm evaluates the presence of keywords inside the content of the website. The simple presence of the keywords in the text of the web page is not enough to consider the content as present. If the content of a web page matches a keyword, our algorithm evaluates also if the match happened in the appropriate context. For example, we can constrain the keyword to be an anchor of the home page pointing to an existing internal page in the same domain.

Quantitative Analysis of the Public Administration Websites

In our case the task of the system was to assess the compliance of the websites of Italian public administration bodies with the guidelines of the government, which require inserting in the home page of the website a minimum set of sections in specified positions.

The definition of the list of keywords was partially driven by the guidelines, which in certain cases are imposed by law. Moreover, we augmented the list with a number of synonyms and related keywords. It is important in this step to avoid as much as possible the use of ambiguous keywords in order to prevent false positives.

We classified the results of the research according to three categories:

1. **Compliant**: the keyword was found in the home page in the format as requested by the guidelines;

2. **Not Compliant**: the keyword was found inside the website, but not in the home page and/or not as prescribed by the guidelines;

3. **Not found**: the keyword was not found in the website.

Results

From the research it emerged that, in 2010, the domain names registered by Italian public bodies (PA) were 25,681, of which 15,279 (see table 1) were assigned to territorial institutions: 947 domains assigned to 20 “Region” bodies (out of 20 Italian regions), 1,022 domains assigned to 109 “Province” bodies (out of 110 Italian provinces) and 13,310 domains assigned to 7,923 “municipality” institutions (out of 8,094 Italian municipalities). The “unique” domain names (excluding aliases, that is the domains associated with the same site) registered by the PA count 23,266.

Table 1 shows in detail the results of the research obtained by means of a crawling operation for the territorial bodies: Region, Province and Municipality. The 15,279 domain names assigned to territorial bodies were subdivided into: 1,949 aliases (domains to which the same website is associated), 1,050 “domains without site” (domains that are not associated with a website), 12,099 valid sites (websites (one or more) associated with a territorial body excluding aliases) and 181 domains “not classifiable” (domains with wrong redirects, pages under construction, etc).

The crawling operation detected sub-domains relative to the domain names contained in the original list, increasing it to 35,236 elements. The complete download of all the analyses pages took about 72 hours. A total of 10,951,256 downloads were carried out, of which 8,951,731 were web pages and 482,582 were redirects to HTTP and other pages. On the whole, within all the downloaded pages there were 319 million hypertext links (319,400,761). The entire crawling activity downloaded data amounting to about 180 GigaBytes.

Table 2, 3, 4 show the results of the evaluation of the quality of websites belonging to the “Region”, “Province” and “Municipality” bodies. This assessment was measured, for each section, in terms of compliance of websites with the guidelines issued by the Italian government. The

results in Table 2 show that, in 2010, all Italian regions had their "Organigram" on their website and this section was compliant with the Government Guidelines (Compliant). 80% of the websites of the regions had a compliant Public Relations Office section whilst 20% were not compliant. The percentage for other sections is also reported.

At the provincial level, as shown in Table 3, 66% of Italian provinces (72 provinces out of 109) have in their website a compliant Organigram section. While for the 30% of province bodies the Organigram is reported but not compliant, only 4% of Italian provinces (4 out of 109 provinces) do not have in their .it websites the Organigram section. For the provinces, in contrast with what was detected in the websites of the regions, the second section which is mostly present in the websites of the provinces, and compliant with the guidelines issued by the Italian government, is the Public competitions" section. In fact, 64% of the provinces have in their websites "Public competitions section compliant, 27% of the provinces have the section, non- compliant while this section is not present in 9% of the provinces. For what concern the provinces, in 2010, 52% of Italian provinces had the Public Relations Office section on their websites and this section was compliant, while 33% of provinces had the Public Relations Office section non-compliant and 15% of provinces did not have that section. As shown in table 2 for regions, also in the case of provinces, only some of them have in their ".it" websites the "Services" section compliant (8%), 58% of provinces have this section, but non-compliant with the guidelines, while for the remaining 34% of provinces this section is not present.

Finally, table 4 shows the results concerning the quality of the websites of "Municipality" bodies. For municipalities, the quality of websites appears to be lower compared to regions and provinces. For example, in the websites of municipalities, the Organigram section is compliant with the guidelines only for 34% (2,706 out of 7,923 municipalities have in their websites a compliant section), 41% have the section but non-compliant, whilst for 25% of municipalities this section is not present. Municipalities, unlike regions and provinces do not have the "Public Relations Office" section on their ".it" websites very often. In fact, only 13% of municipalities have this section compliant, in 24% is non-compliant, while for the remaining 63% the section is not present. For municipalities, the section that is more compliant with the guidelines issued by the Government is the "certified e-mail" section. 42% of municipalities have a compliant section, 18% is not compliant, while for 40% it is not present. Moreover, as reported in table 2 and 3 for regions and provinces, also for municipalities the "Administrative Procedures" section is almost always absent in their .it websites. Only 13% of monitored municipalities have this section in their websites, but it does not comply with the guidelines, while for 87% of municipality websites this section is not present.

In 2011, following on from the analysis carried out in 2010, an update concerning the presence of the "Transparency"[2] and "Publication of legal notices" sections in ".it" websites of Italian municipalities was carried out. According to this study, in July 2011,

[2] Transparency, intended as "total accessibility", consists in the publication on corporate websites of the Italian Public Administration of all the information concerning organization, curricula, salaries, absence and presence rates of personnel and results of the monitoring and evaluation activity carried out by the competent bodies, are just some of the essential elements to foster the diffusion of monitoring strategies in accordance with principles of good conduct and impartiality. Transparency represents, in fact, a prerequisite of the services supplied by the Italian Public Administration in accordance with art. 117, second paragraph, letter m) of the Constitution. With the Legislative Decree No.150 October 27, 2009, the Legislator has defined a set of mandatory contents that Italian Public Administrations have the duty to publish in a relevant section on their corporate websites. In particular, CiVIT Deliberation No. 105/2010 "Guidelines for the elaboration of the triennial programme for transparency and integrity" specifies, among other things, the contents that must be published on the corporate site and the modalities of publication in order to facilitate retrieval and use by citizens.

there were 7,913 municipality bodies that registered at least one ".it" domain name and in most cases, at least one domain name per body was associated with a website. Monitoring results highlights that, in July 2011, the amount of municipality bodies whose website provides the "Transparency" section exceeds 77% (6,119 out of 7,913 analyses), against 54% (4,300 out of 7,923 municipalities analyses) of the survey conducted in 2010 (see table 4). Figure 3 shows the number of municipality bodies aggregated by region, which have within their websites or immediately visible on the homepage, the "Transparency" section clearly identifiable. Figure 3.1 shows the percentages of municipality bodies which have the transparency section for each Region. In 2011, in almost all the regions, more than 50% of municipalities have the transparency section on their ".it" web sites, while this section appears only for 12.81% in the websites of the municipalities of the region Trentino Alto Adige.

Figure 4 shows, however, a special section devoted to the verification of the state of law enforcement containing dispositions aimed at the elimination of waste relating to the maintenance of paper documents, which came into force in Italy, on January 1 2011 (Article 32 of Italian Law No. 69/2009). As a result, paper publications no longer have legal status: all Italian Public Administrations are obliged to publish on their website - or on those of other similar administrations or associations – all the news and administrative acts that require legal publicity (bids for tenders, building licenses, resolutions of the town Board and Council, list of the beneficiaries of economic provisions, wedding bands, etc..). Compared to the previous analysis carried out in 2010 (which had shown that 61%) (4,796 out of 7,923, see table 4) of the municipalities had set up on their institutional website an online section of 'Publication of legal notices' or legal publicity the results of the survey conducted in July 2011 show that 7,718 municipalities with a ".it" website out of 7,913 (equal to 97.54% of the municipalities taken into account) have the Publication of legal notices section and are, therefore in line with the Guidelines provisions. Figure 4 shows, the number of municipalities, aggregated by region, having a ".it" website and on-line 'Publication of legal notices' section. Figure 4.1 shows the percentage of Municipality bodies which have the 'Publication of legal notices' section in their institutional websites. Municipality bodies were subdivided by region. As shown in figure 4.1, almost all municipalities of the various regions have the Publication of legal notices section in their .it website with a percentage that ranges from the 88% of Trentino Alto Adige to the 100% of Valle d'Aosta.

Table 1 - Number of domain names registered by Region, Province and Municipality bodies

	Number in Italy	Territorial bodies analyzed	Total domains	Aliases	Domains without site	Valid sites(not aliases)	Not classifiable
REGION	20	20	947	336	219	347	45
PROVINCE	110	109	1022	263	104	632	23
MUNICIPALITY	8,094	7,923	13,310	1,350	727	11,120	113
TOTAL			15,279	1,949	1,050	12,099	181

Table 2 - Classification of the sections present in websites of the Italian Regions			
Section	Not Found	Compliant	Not Compliant
Organigram	0	20	0
URP (Public Relations Office)	0	16	4
Transparency	5	3	12
Administrative Procedures	4	1	15
Calls for tender	2	6	12
Public competitions	0	14	6
Services	6	4	10
Publication of legal notices	10	2	8
PEC (certified e-mail)	2	13	5

Table 3 - Classification of the sections present in websites of the Italian Provinces			
Section	Not Found	Compliant	Not compliant
Organigram	4	72	33
URP (Public Relations Office)	16	57	36
Transparency	27	32	50
Administrative Procedures	52	3	54
Calls for tender	34	13	62
Public competitions	10	70	29
Services	37	9	63
Publication of legal notices	39	37	33
PEC (certified e-mail)	21	63	25

Table 4 - Classification of the sections present in websites of the Italian Municipalities			
Section	Not Found	Compliant	Not compliant
Organigram	1958	2706	3259
URP (Public Relations Office)	5008	998	1917
Transparency	3623	1454	2846
Administrative Procedures	6858	22	1043
Calls for tender	5234	477	2212
Public competitions	3326	2470	2127
Services	5430	351	2142
Publication of legal notices	3127	2849	1947
PEC (certified e-mail)	3205	3323	1395

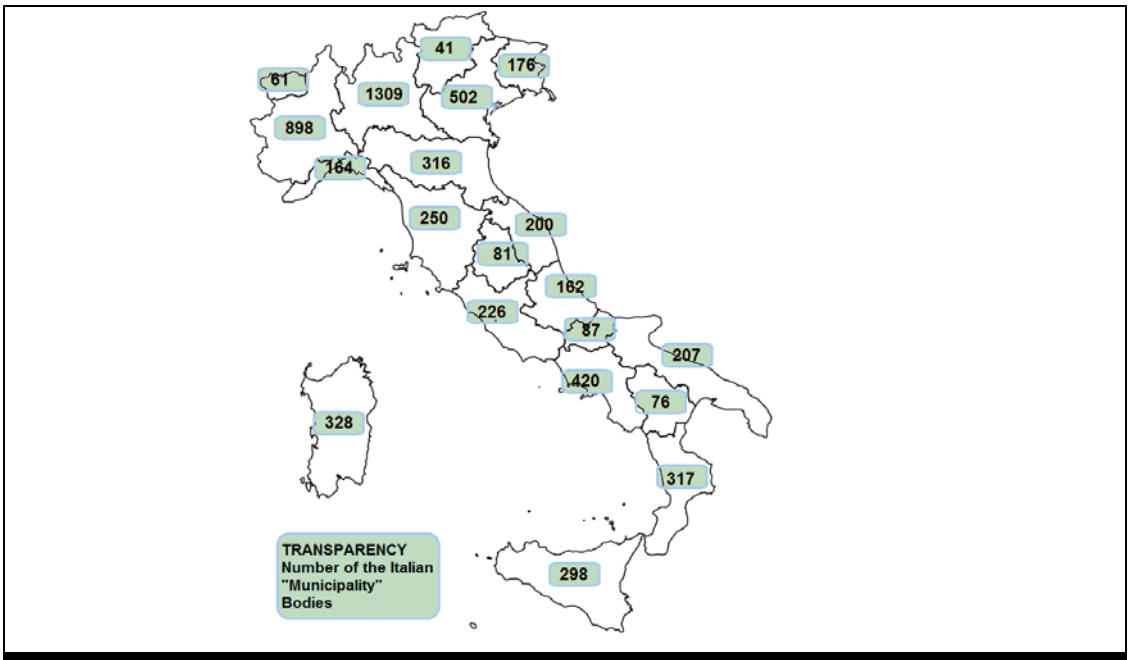


Figure 3 - Municipalities having the transparency section in their ".it" websites

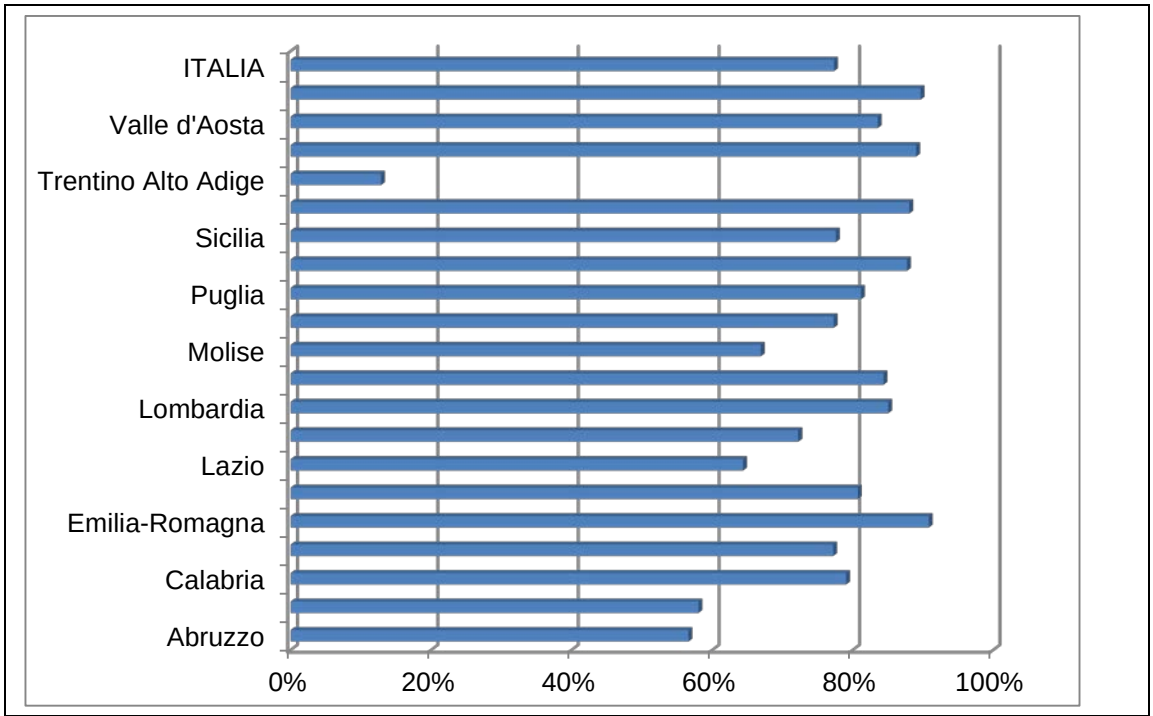


Figure 3.1. Percentage of municipality bodies which have the transparency section

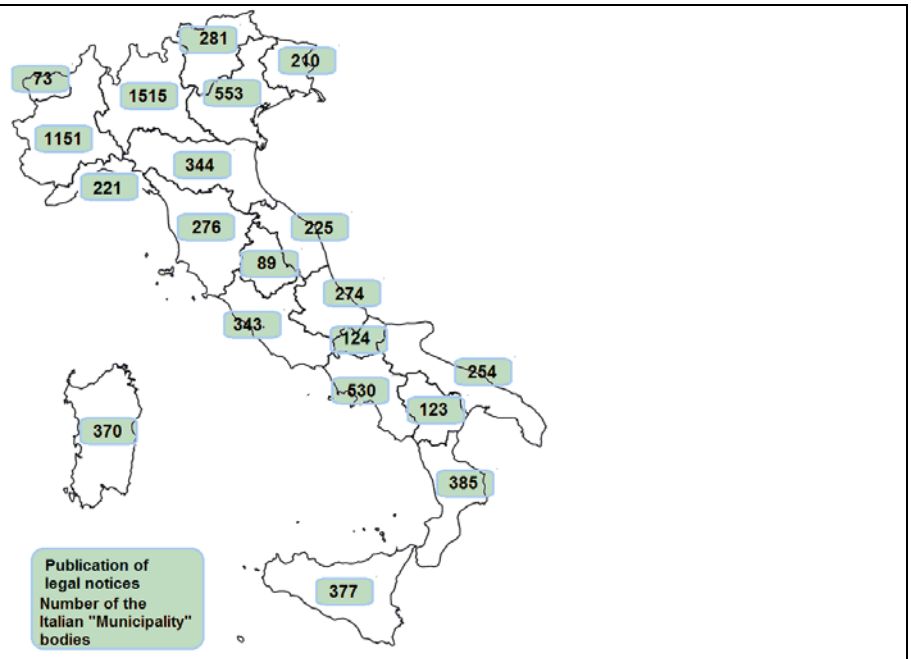


Figure 4 – Publication of legal notices. Municipalities aggregated per Region

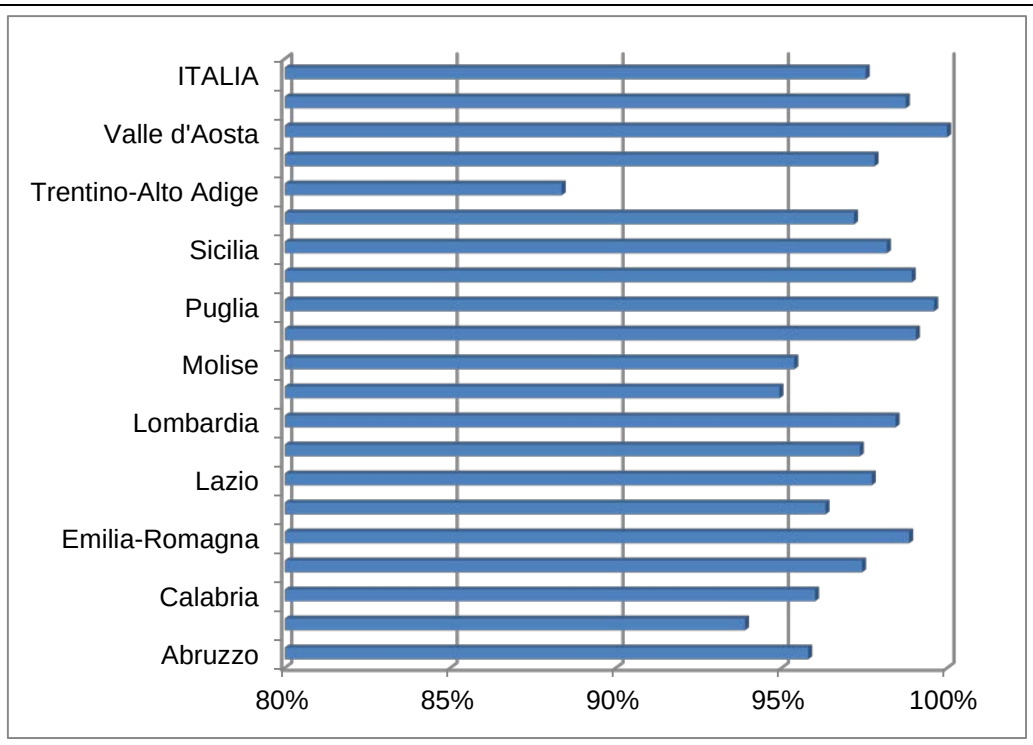


Figure 4.1 - Publication of legal notices section. Percentage of Municipalities

Discussion

This paper introduces, for the first time, an experimental approach for the evaluation of the quality of PA websites. This approach is based on the extraction of the list of all domain names registered by local government bodies ("Provinces", "Regions", and "Municipalities") taking as reference the case of Italy. The availability of all domain names and all of the websites of the provinces, regions and municipalities, allowed us to evaluate in a fully automatic way the quality of their websites. This evaluation was carried out with the use of a web-crawler that allowed us to examine, with extreme accuracy the contents of the PA web pages. Compared to traditional approaches such as ad hoc surveys, this framework has the advantage of getting real and incontestable results, in a short time, avoiding any distortion in the final evaluations. In addition, with this approach it is possible to monitor the quality of websites at all times. In practice, our framework allowed us to verify in a fully automatic mode, if the sites of the main Italian local PA are in line with the current legislation. These results are fundamental to the central and local governments so that they have a continuous monitoring of their websites, without the use of questionnaires that would limit the sample to be examined, with great expenditure of resources and with the onset of possible distortions of data.

Our results showed that, in accordance with literature (Moon, 2002), in 2010, the Italian PAs with smaller size (the municipalities) have a lower quality of their websites compared to PAs with larger size (regions and provinces). In fact, in 2010, some sections prescribed by the regulations are not always found to be present in the websites of Italian municipalities, whereas the same sections were found to be present in the websites of regions and provinces. However, if we examine the results of 2011, the municipalities seem to have adapted to the above requirements. By the temporal analysis of the data obtained in this research, it is noted that in Italy the largest public administrations are more innovative than smaller ones, as the former have adapted more quickly to the legislation than the latter.

It is interesting to note that our framework can also be used by other institutions in other countries who want to study the quality of e-government in the public sector. This approach, based on automatic procedures allows studying carefully every single website, with a possible comparison both temporal and spatial, without resorting to typical questionnaires as reported in the literature (e.g., Torres et al., 2005 and Scavo, 2003).

Conclusions

This paper presents an experimental methodology of assessment for the quality of the websites of Public Administration bodies and measurement of the level of use and efficiency of their services. This methodology was tested by extracting the list of domain names registered by local government bodies. These data were taken from the database of the ccTLD ".it" Registry, managed by the Institute of Informatics and Telematics of the CNR (National Research Council) of Pisa, Italy. Access to this database enabled us to have a comprehensive sample of data and therefore avoid incongruences and inaccuracies in the final results.

For the quantitative assessment of the information content of the PA websites, compliancy of the institutional websites with government guidelines were verified. The guidelines lay down specific criteria for some information sections which must be present on the home page or immediately available by means of hypertext links from the homepage. Furthermore, for these sections there is specified a list of possible alternative texts that must be contained within the anchor of the hypertext link. In accordance with these guidelines, for each website section, the local government websites have been classified in three categories, compliant (strict adherence to

the guidelines), not compliant (within the whole text of the site there is just one keyword among those specified in the guidelines, even if it is not in the position or format indicated by the guidelines), not found (the keywords searched for do not appear within the site, or the structure and technology used in the construction of the website does not permit an automatic analysis of its content, at the present state of the procedures used for the analysis).

This research also showed that there is a difference in the quality of the websites of the bodies taken into account (Region, Province and Municipality): as the size of the bodies decrease (from Region to Municipality), the quality of their websites also decreases. It was in fact demonstrated, for at least some sections that Region and Province bodies have in their websites the sections foreseen by Italian Law, while in the websites of Municipality bodies these sections are not present or are not compliant with the guidelines issued by the Italian government.

This approach, aimed at measuring and analyzing the capacity of the Italian PA bodies to activate and manage information and services by means of communication via the Internet and web channels, could be a useful point of reference for other similar government agencies. These institutions can apply this technique for the evaluation of their websites by defining a concise quality indicator, measured in terms of coherence or degree of compliancy with a minimum set of mandatory requisites and/or contents.

Acknowledgements

This work was supported by the Department for Digitalization of the Public Administration and the Technological Innovation of the Presidency of the Council of Ministers and the Institute for Informatics and Telematics of the Italian National Research Council (IIT-CNR).

References

Alshehri, M., Drew, S., Alhussain, T. and Alghamdi, R. (2012). "The Effects of Website Quality on Adoption of E-Government Service: An Empirical Study Applying UTAUT Model Using SEM", *Proceedings of the 23rd Australasian Conference On Information System.*

Boldi, P., Santini, M. and Vigna, S. (2004). "Do your worst to make the best: Paradoxical effects in pagerank incremental computations", *Proceedings of the Third International Workshop on Algorithms and Models for the Web-Graph*.

Castillo, C. (2004). "Effective Web Crawling", *Ph.D. thesis, University of Chile*.

Diez-Picazo, G.F.(1999). "An Analysis of International Internet Diffusion", *Ph.D. Thesis, MIT.*

Edwards, J., McCurley, K. and Tomlin, J. (2001). "An adaptive model for optimizing performance of an incremental web crawler", *Proceedings of the 10th international conference on World Wide Web*.

Fong, S. and Meng, Ho Si. (2009). "A web-based performance monitoring system for e-government services", *Proceedings of the 3rd international conference on Theory and practice of electronic governance*.

Felicioli, C., Geraci, F. and Pellegrini, M. (2011). "Medium sized crawling made fast and easy through Lumbricus webis", *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics*.

Geraci, F., and Maggini, M. (2011). "A Multi-sequence Alignment Algorithm for Web Template Detection", *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, KDIR.*

Marin, M., Paredes, R. and Bonacic, C. (2008). "High-performance priority

queues for parallel crawlers", *Proceeding of the 10th workshop on Web information and data management, ACM*.

Martinelli, M., Serrecchia, I. and Serrecchia, M. (2006). "Analysis of the Internet diffusion in the non-profit sector: the social digital divide in Italy", *Scientometrics,* 66(1), pp. 155-170.

Moon, M. J. (2002)."The Evolution of E-Government among Municipalities: Rhetoric or Reality", *Public Administration Review*, 62(4), pp. 424-433.

Olston, C. and Najork, M. (2010). "Web crawling", *Foundations and Trends in Information Retrieval*, 4(3), pp. 175-246.

Scavo, C. (2003). "World Wide Web Site Design and Use in Public Management", in Garson, G.D. (ed.), *Public Information Technology: Policy and Management Issues*, Idea Group Publishing: Hershey, PA.

Sousa, J.M.E. and Lopez, M.V.W. (2007). "Analyzing Municipal e-Government in Peru", *Proceedings of the 9th International Conference on Social Implications of Computers in Developing Countries.*

Thelwall, M. and Stuart, D. (2006). "Web crawling ethics revisited: Cost, privacy, and denial of service". *J. Am. Soc. Inf. Sci. Technol.*, 57(13), pp. 1771-1779.

Torres, L., Pina, V. and Royo, S. (2005). "E-government and the transformation of public administrations in EU countries: Beyond NPM or just a second wave of reforms?", *Online Information Review*, 29(5), pp.531 – 553.

Wang, L. and Bretschneider, S. (2005). "Evaluating Web-based e-government services with a citizen-centric approach", *Proceedings of the 38th Hawaii International Conference on System Sciences*, *IEEE Computer Society.*

About Authors

Filippo Geraci was born in Palermo in Sept. 20, 1977. He is researcher at the Institute for Informatics and Telematics since 2009. From 2005 to 2009 he was research assistant in the same institute. Since 2010 he holds the chair of "Information systems for business management" at the Information Engineering Department at the Siena University. In April 2008 he has completed the Ph.D. in "Information Engineering" at the Siena University. Since 2010 he cooperates on behalf of the IIT-CNR with the Department for Technological Innovation (DIT) and he contributed to a study published in the Rapporto egov Italia 2010". His research activity covers the fields of Information Retrieval and Bio-Informatics.

Maurizio Martinelli, born on October 15th 1964, received his laurea degree in Computer Science at the University of Pisa. He joined the Italian National Research Council (CNR) on 1992. From 1992 to 1994 he was the technical coordinator of the Italian X.500 Directory Service, participating to the *DIR-ITA* national project, the Italian branch of the VALUE Subprogramme II, an European project financed by D.G. XIII whose aim was the design, implementation and diffusion of the X.500 Directory Service technology in Europe. From 1995 to 1999 he was the technical coordinator of the Italian/Egyptian Cooperation project *Mubarak City*, financed by the Italian Ministry of Foreign Affairs and whose aim was the computerization of the Informatics Research Institute (IRI) located in New Borg El Arab - 40 km south of Alexandria -, and the technical training of the IRI staff both in Egypt and in Italy. From 1998 he is the technical manager of the .it ccTLD and head of the *Systems and Development Department* of the .it Registry, whose aim is the design and the implementation of advanced technologies for the management of new generation domain names Registries. From 2002, he is also the head of the *Internet Services and Technological Development Department* of the Institute of Informatics and Telematics of the Italian National Research Council (IIT-CNR), formed by several laboratories and whose purpose is the planning, designing and development of innovative telematics applications for the Institute and, more in general, for the CNR, the Italian Public Administration and the private sector.

Marco Pellegrini, born on June 11th 1961, received his laurea degree (magna cum laude) in Electronic Engineering at Polytechnic of Milan in 1986. In 1991 he was awarded a Ph.D. in Computer Science by the New York University and a postdoctoral fellowship by the International Computer Science Institute in Berkeley. From 1991 to 1995 he was lecturer at the Department of Computer Science of King's College, London. In 1995 he joined C.N.R. of Italy where he is Senior Scientist as of 1998. His research interests are in Bioinformatics, Data mining, Computational Geometry and Analysis of Algorithms with more than sixty research papers in international journals and conferences. Currently he coordinates focus groups in IIT on algorithmics for biology (BioAlgo), and on web data mining (WebAlgo). He coordinates activities of IIT for the 7FP Network of Excellence VPH (Virtual Physiological Human) in partnership with ERCIM. He is scientific leader for IIT of the joint IIT-IFC laboratory LISM (Laboratory for Integrative Systems Medicine) and of the IIT unit within the MIUR funded flag project InterOmics.

Michela Serrecchia is research associate at the Institute for Informatics and Telematics of Italian National Research Council (IIT-CNR) since 2004. She received the laurea degree in Business Economics from Pisa University in 2004 and the Ph.D. degree in Applied Statistics from Florence University in 2011. Her research interests are mainly devoted to analysis of Internet Diffusion and Digital Divide. She is author of several scientific papers in international journals, book chapters and conferences.