Automatic Identification and Preservation of R&D Websites

Daniel Bicho and Daniel Gomes

Foundation for Science and Technology: Arquivo.pt
Av. do Brasil, 101
1700-066 Lisboa, Portugal
(daniel.bicho,daniel.gomes)@fccn.pt

Abstract. Research and Development (R&D) websites often provide valuable and unique information such as software used in experiments, test data sets, gray literature, news or dissemination materials. However, these sites frequently become inactive after the project ends. For instance, only 7% of the project URLs for the FP4 work programme (1994-1998) were still active in 2015. This study describes a pragmatic methodology that enables the automatic identification and preservation of R&D project websites. It combines open data sets with free search services so that it can be immediately applied even in contexts with very limited resources available. The "CORDIS EU research projects under FP7 dataset" provides information about R&D projects funded by the European Union during the FP7 work programme. It is publicly available at the European Union Open Data Portal. However, this dataset is incomplete regarding the project URL information. We applied our proposed methodology to the FP7 dataset and improved the completeness of the FP7 dataset by 86.6% regarding the project URLs information. Using these 20 429 new project URLs as starting point, we collected and preserved 10 449 947 Web files, fulfilling a total of 1.4 TB of information related to R&D activities. All the outputs from this study are publicly available [17], including the CORDIS dataset updated with our newly found project URLs.

1 Introduction

Most current Research & Development (R&D) projects rely on their websites to publish valuable information about their activities and achievements. However, these sites quickly vanish after the project funding ends. During the funding work programme FP7 the European Union invested a total of 59 107 million EUROS on R&D projects. Scientific outputs from this significant investment were disseminated online through R&D project websites. Moreover, part of the funding was invested in the development of the project websites themselves. However, these websites and the information they provide typically disappear a few years after the end of the projects. Websites of R&D projects must be preserved because:

- They publish valuable scientific outputs;
- They are aggregators of scientific outputs related to a given theme because the R&D projects are typically funded in response to a call on proposals to solve specific societal or scientific problems;

– They are not being exhaustively preserved by any institution;
– They are highly transient, typically vanishing shortly after the project funding ends;
– They constitute a trans-national, multi-lingual and cross-field set of historical web data for researchers (e.g. social scientists).

The constant deactivation of websites that publish and disseminate the scientific outputs originated from R&D projects causes a permanent loss of valuable information to Human knowledge from a societal and scientific perspective. Web archiving provides a solution to this problem. Web archives can preserve this valuable information. Moreover, funding management datasets can be enriched with references of the preserved versions of the project websites that disappeared from the live-Web. However, websites that publish information related to R&D projects must be firstly identified so that web archives can preserve them.

There has been a growing effort of the European Union, and governments in general, to improve transparency by providing open data about their activities and outputs of the granted fundings. The European Union Open Data Portal [8] is an example of this effort. It conveys information about European Union funded projects such as the project name, start and end dates, subject, budget or project URL. Almost all this information is persistent and usable through time after the project or funding instruments end. The exception is the project URL. As websites typically disappear a few years after their creation [33], the R&D management databases available at The European Union Open Data Portal, such as the datasets of the CORDIS EU research projects, suffer degradation by referencing complementary online resources that became unavailable and were not systematically preserved neither by the funder nor the funded entities. Moreover, the CORDIS EU research project datasets have incomplete information regarding the projects URLs. From a total of 25 608 project entries, only 2 092 had the project URL field filled. Thus, about 92% of project websites could not be identified and therefore their preservation was challenged.

The Foundation for Science and Technology (FCT) [10] is the official Portuguese institution that manages research funding and e-infrastructures. Arquivo.pt [2] - the Portuguese Web Archive is one of the research infrastructures managed by FCT and its main objective is to preserve web material to support research activities. Hence, the websites of R&D projects are priority targets to be preserved. The objective of our work was to study techniques to automatically identify and preserve R&D project websites funded by the European Union based on existing free tools and public data sets so that they can be directly applied by most organizations and information science professionals, without requiring the intervention of computer scientists, or demanding computing resources (e.g. servers, bandwidth, disk space). The main contributions of this work are:

– Quantitative measurements about the ephemera of EU-funded project websites and their preservation by web archives;
– A test collection and methodology to evaluate heuristics to automatically identify R&D project websites;
– A comparative analysis between heuristics to automatically identify URLs of R&D projects using free search services and publicly available information datasets;

– A list of web addresses of existing R&D project sites that can be used by web archives to preserve these sites or by management institutions to complement their datasets.

We believe that the results described here can be immediately applied to bootstrap the preservation of EU-funded project websites and minimize the loss of the valuable information they convey as has been occurring for the past 22 years.

2 Related Work

The vastness of the web represents a big challenge with regard to preservation activities. Since it's practically impossible to preserve every web content, the question remains: "how much of the web is archived? [22]". The problem of link rot is a serious and prevalent problem that jeopardizes the credibility and quality of scientific literature that increasingly references complementary online resources essential to enable the reproducibility of the published scientific results (e.g. experimental data). A study about the decay and half-life period of online citations cited in open access journals showed that 24.58% of articles had online citations and 30.56% of them were not accessible [44]. The half-life of online citations was computed to be approximately 11.5 and 9.07 years in Science and Social science journal articles respectively. However, the link rot problem in scientific publications is not a problem of open access journals. The unavailability of online supplementary scientific information was also observed across articles published in major journals [30,32]. The problem of link rot is cross-field and has been scientifically reported over time. For instance, it was observed among scientific publications in the fields of Computer Science in 2003 [47], Information Science [48] in 2011 or Agriculture in 2013 [46]. We believe that many of the link rot citations reference resources published on project websites that meanwhile became unavailable. Preserving these sites would significantly contribute to maintain the quality of scientific literature.

Since the early days of the web, several studies addressed the problem of identifying relevant web resources. Focused crawling approaches try to identify valuable information about a specific topic [27]. ARCOMEM - From collect-all archives to community memories was a EU-funded research project conducted between 2011 and 2013 that aimed to study automatic techniques to identify and preserve relevant information regarding given topics specially from social media. Ironically, the project website is no longer available and could only be found in publicly available web archive [24]. ARCOMEM studied, for instance, how to perform intelligent and adaptive crawling of web applications for web archiving [31] or how to exploit the social and semantic web for guided web archiving [42]. However, implementation of such approaches is too complex and entails a significant amount of resources, requering powerful crawlers and bandwidth resources to harvest the web looking for relevant resources. The process can be optimized but considering the dimensions of web data, it is still too demanding to be implementable by most Cultural Heritage Organizations. web services, such as live-web search engines, have already crawled and processed large amounts of web data, and provide search services to explore it. Bing Web Search API [3] and Google Custom Search API [12] are examples of commercial APIs that can be used to explore those web data. However, these services limit the number of queries per user based on the subscribed plan. Contrarily, non-commercial APIs like Faroo [9] don't have limitations on the number of queries a user can perform, but the search results tend to be worse due to the relatively low amount of web data indexed.

Therefore, alternative approaches that explore existing services and resources to identify and preserve relevant web content have been researched. Martin Klein and Michael Nelson proposed methods to rediscover missing web pages automatically through the *web Infrastructure* [35]. In their study they have *a priori* information about the original URL which they used it to build several heuristics to rediscover the missing web pages. Shipman et al. used page titles to rediscover lost web pages referenced on the DMOZ web directory by using the Yahoo search engine [45].

Websites containing information regarding European Union fundings and R&D projects are frequently referenced by names under the .EU domain. There is no entity in charge of preserving the general content published under the .EU domain. The strategy adopted by memory institutions has been to preserve the web through the delegation of the responsibility to each national institution which leaves the content published under the .EU domain orphan regarding its preservation. Nonetheless, the Historical Archives of the European Union (HAEU), in cooperation with the EU Inter-institutional Web Preservation Working Group coordinated by the EU Office of Publications, has started a web archiving pilot project in late 2013 concerning the websites of EU institutions and bodies. They performed four complete crawls of 19 EU Institutional and Bodies websites in 2014 and extended this to include 50 EU Agencies in 2015 [21]. Arquivo.pt performed a first exploratory crawl of the .EU domain to gain insight into the preservation of the content published under this domain [25]. The initial idea was that the "brute-force" approach of preserving the .EU websites in general would also include most R&D projects websites hosted on this domain. However, the obtained results showed that this approach was too demanding for the resources we had available. Therefore, we decided to adopt a more selective approach. By combining open data sets and free search services, we have established a pragmatic framework that enables the automatic identification and preservation of R&D project URLs in contexts with very limited resources available.

3 Ephemera of R&D websites

Everyday, more information is published on the web, from a simple blog post opinion to a research project funded by the European Union. However, the web is ephemeral. Only 20% of web pages remain unchanged after one year, which points towards a massive loss of information [40]. We performed an experiment to measure the ephemera of research websites funded by the European Union work programmes from FP4 (1994-1998) to FP7 (2007-2013). On the 27th November 2015, we tested the available projects URLs for each funding work programme (FP4 [4], FP5 [5], FP6 [6] and FP7 [7]), checking how many still referenced relevant content. The datasets containing the projects URLs was obtained from the European Union Open Data Portal datasets [8]. A comparison was made using the *title* on the datasets and the project URL content to test if each project URL was still referencing relevant content. The relevance criterion applied was that if at least half the words with 4 or more characters presented on the *title* were found on the content referenced by the project URL, the content was considered to be relevant. This method was applied on all work programmes with exception of FP7 that was humanly validated to build the test collection described in Section 4.

Table 1. Project URLs from the CORDIS dataset referencing relevant content distributed per work programme validated in 27 November, 2015.

	Nr. project URLs	Nr. project URLs with relevant content	% project URLS relevant content
FP4 (1994-1998)	853	58	7%
FP5 (1998-2002)	2 717	322	12%
FP6 (2002-2006)	2 401	715	30%
FP7 (2007-2013)	2 092	1 370	65%

The results presented on Table 1 show that 65% of the URLs of R&D projects funded by FP7 program were still available and referenced relevant content. A counterexample of a R&D project URL, presented on the FP7 dataset, that now references irrelevant content is `www.oysterecover.eu`. This URL is associated to the OYSTERECOVER project that studied scientific bases and technical procedures to recover the European oyster production, and now references a shopping website. The percentage of active and relevant project URLs decreased for older work programmes, reaching a percentage of only 7% for the FP4 work programme (1994-1998).

3.1 Preservation Coverage and Distribution

Our previous results showed that a significant percentage of project URLs is no longer available on the live-web and therefore its content may have been potentially lost forever. However, there are several web Archiving initiatives working to preserve the web as exhaustively as possible. Many of them focus on the preservation of each respective country web domain, with some exceptions like the US-based Internet Archive [14], a non-profit initiative that acts with a global scope.

We conducted an experiment to measure if the available project URLs referenced on the incomplete CORDIS datasets were preserved by web archives. For this purpose, we verified if at least one web-archived version of the referenced project URLs could be found by using the Time Travel Service [20,23]. This service acts as gateway to query for archived versions of a web resource (Memento) across multiple publicly available web archives using the HTTP Memento Framework [28]. For each project URL, we queried the Time Travel Service for its *timemap* which provides a list of corresponding archived versions. If a project URL had an archived version between the time range of the corresponding work programme, we considered that the project URL had a valid archived version. The results of this experiment are presented on Table 2. It shows that 1 593 of the 2 092 FP7 project URLs have an archived version between 2007 and 2013, meaning that 76.1% of these projects URLs have an web-archived version. However, the amount of project URLs preserved decreases for the older work programmes, only 38.2% of the FP6 project URLs had a web-archived version, and 43.6% for FP4 project URLs.

Table 2. Projects URLs on EU CORDIS datasets with a web-archived version.

	Nr. project URLs	Nr. project URLs with an archived version	% project URLs with an archived version
FP4 (1994-1998)	853	372	43.6%
FP5 (1998-2002)	2 717	1 661	61.1%
FP6 (2002-2006)	2 401	918	38.2%
FP7 (2007-2013)	2 092	1 593	76.1%

Table 3. Distribution of projects URL archived versions per web archive.

Time Gates	% FP4	% FP5	% FP6	% FP7	% Average
web.archive.org [14]	43.61	60.91	37.90	76.0	54.61
web.archive.bibalex.org [13]	12.54	22.56	21.90	0.72	14.43
webarchive.loc.gov [15]	0	1.80	0.58	0.43	2.81
webarchive.nationalarchives.gov.uk [49]	0.12	0.22	0	0.57	0.91
arquivo.pt [2]	0.47	0.55	0.24	0.67	0.48
wayback.archive-it.org [1]	0	0.04	0	0.81	0.21
wayback.vefsafn.is [39]	0.35	0.03	0.08	0.23	0.17
webarchive.parliament.uk [49]	0	0	0	0.19	0.05
www.webarchive.org.uk [26]	0	0	0	0.19	0.05
www.padi.cat [50]	0	0	0	0.04	0.01
collection.europarchive.org [34]	0	0	0	0.04	0.01

Table 3 shows the distribution of the project URLs archived versions across web archives. For each project URL we counted how many web archives have a valid archived version. Most of the project URL archived versions are retrieved from `web.archive.org`, the time gate of the Internet Archive, with 76% preservation coverage of the FP7 project URLs followed by web.archive.bibalex.org with only 0.81% of the FP7 project URL preserved. This results show that EU-funded project URLs were mainly preserved by the US-based Internet Archive.

4 Evaluation methodology

This section describes the evaluation methodology used to compare the performance of the heuristics tested to automatically identify R&D projects websites. We present here test collection developed as well as the relevance criterion adopted.

4.1 Test Collection

A ground-truth is required to evaluate the performance of the proposed heuristics. We developed a test collection based on the FP7 dataset [7]. The objective was to build a list of carefully validated pairs of projects and corresponding project URLs. The CORDIS dataset contains several information fields about each funded project such as *acronym* (project acronym), *title* (description of the project) and *projectUrl* (URL for the project site or page). However, for most projects the URL was missing. Thus, we removed all the projects that had the *projectUrl* field blank, ending up with a list of 2 092 entries with *projectUrl* filled. Then, the following data cleansing steps were applied to the dataset:

1. Removed non-existent URLs or invalid URLs (return codes that are not 200s or 300s);
2. Followed all redirects and updated the *projectUrl* field with the target URL;
3. Removed non alphanumeric characters from the title fields;
4. Left and right trim each column and removal of multiple white spaces.

The dataset resulted in a list of 1 596 entries. However, this list was still not ready to be used as a test collection because there were project URLs referencing online content no longer related to the project. For example, some URLs projects referenced registrar sites, shopping sites or Chinese sites that became the new owners of the the domain names. A human validation was performed to overcome this situation, deleting entries where the *project URLs* were no longer related to the project. With this manual validation the test collection ended up with a list of 1 370 project entries with valid project URLs.

The search engines API have some limitations regarding how many queries can be made. For example, Google Custom Search Engine has a limitation of 100 search queries per day, and Bing Web Search API has a limitation of 5 000 queries a month for free usage. These limitations slowed the heuristics evaluation experiments to identify the R&D projects URLs. For our test collection of 1 370 entries, it would only be possible to experiment 3 heuristics each month. To be able to test several heuristics in a reasonable time, a smaller collection was built from the full test collection. This smaller test collection comprised a random sample of 300 entries from the base test collection, with a confidence level of 95% and a 5% margin of error [18].

4.2 Relevance Criterion

Ideally, the project URLs identified through an heuristic should match the project URL on the test collection. However, a strict string comparison to match URLs raises problems. For instance, it would not detect URLs with different domains but the same content like `www.lipididiet.progressima.eu/` and `lipididiet.eu/`, nor the absence

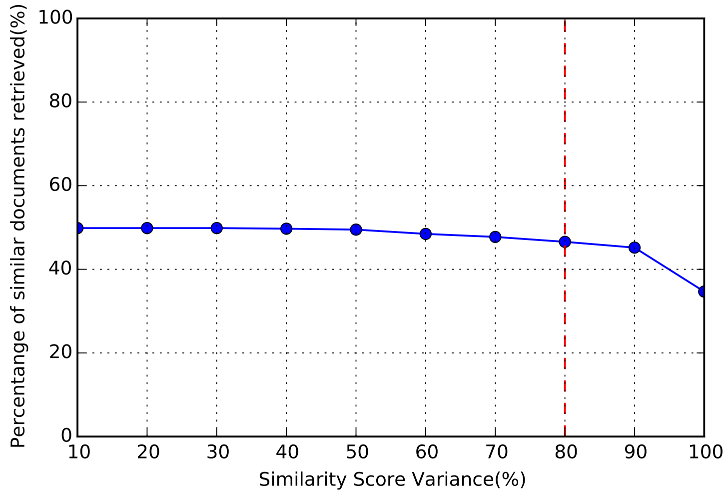

Fig. 1. Fuzzy hashing threshold applied to identify relevant project URLs.

or presence of *www* hostname, `www.hleg.de` and `hleg.de`. Another problematic situation would be different paths names to the same content such as `www.tacmon.eu/new/` and `www.tacmon.eu/`. Thus, we adopted an automatic content comparison approach by using hashing techniques instead of URL comparisons. However, the use of strict hashing techniques like MD5 [43] or SHA-1 [29] to verify if the content referenced by the project URL is relevant also present limitations. Project URLs that reference hidden dynamic content, for instance a simple blank space or a hidden HTML section inserted dynamically would result in totally different hash codes, leading wrongly to the conclusion that the content is not relevant. For this reason, we decided to apply a fuzzy hashing technique [36]. This technique allows us to overcome the previous problems since it generates an hash code proportional to the level of difference between contents. Noteworthy, the similarity threshold cannot be too high (e.g. 100%) because it would suffer from the limitations of strict hashing techniques causing the exclusion of relevant project URLs. On the other hand, the threshold cannot be too low because it would include irrelevant results. The similarity threshold was determined by gradually decreasing the similarity threshold and counting the percentage of relevant results retrieved. Figure 1 shows the percentage of relevant project URLs identified as the fuzzy hashing threshold value increased. We adopted a threshold of 80% for the matching score because the number of similar documents retrieved did not significantly varied below this value and a high percentage of similarity was found. Therefore, we defined that a project URL provided by a search engine is a relevant result for a given project if its content matches the content of the project URL defined on the test collection with a similarity level of at least 80%.

For each heuristic it was measured how well it performed on retrieving the project URL for each project entry on the test collection. An example of a relevant retrieval is when we apply a heuristic to query a search engine about a given project and it returns the URL of the home page of the website with a similarity score of more than 80% in comparison to the test collection project URL content.

5 Heuristics to automatically identify R&D project URLs

Several heuristics to automatically identify project URLs on the live-web were tested. The main idea of these heuristics is to use search engines retrieval capabilities to identify URLS of research project websites.

5.1 +Acronym +Title

This heuristic consists on querying the search engines using the Acronym and Title fields of the FP7 dataset, despite its name provides a textual description of the project. An example of a query submitted to a search engine using this heuristics is: ”IMPACT Impact Measurement and Performance Analysis of CSR”.

5.2 +Acronym +Title -Cordis

This heuristic consists on querying the search engine using the Acronym and Title fields but excluding the results from site `cordis.europa.eu`. The rational behind this exclusion is that search engine results can be biased towards results hosted on the CORDIS site since the query terms used were obtained from the CORDIS datasets. An example of a query submitted to a search engine using this heuristics is: ”IMPACT Impact Measurement and Performance Analysis of CSR -site:cordis.europa.eu”.

5.3 +Acronym +Title -Cordis -EC

This heuristic is the same as the **+Acronym +Title -Cordis** but also excludes the site `ec.europa.eu`. An example of a query submitted to a search engine using this heuristics is: ”IMPACT Impact Measurement and Performance Analysis of CSR -site:cordis.europa.eu -site:ec.europa.eu”.

5.4 +Acronym +Title -Cordis -EC +CommonTerms

This heuristic aims at improving the results returned by search engines through the inclusion of additional query terms commonly used on the content referenced by existing project URLs. The most frequent words extracted from the test collection projects websites content, were identified and then the queries were built by adding these common terms to the query issued to the search engine.

$$\mathbf{v_m} = sort\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_{d_i}\right) \quad (1)$$

Table 4. Top 10 most common terms in the web content referenced by project URLs validated to build the test collection.

Position	Term	Average TF-IDF
1	project	0.048
2	research	0.023
3	european	0.021
4	home	0.017
5	news	0.017
6	eu	0.015
7	new	0.014
8	2015	0.014
9	read	0.014
10	partners	0.014

The method to compute these terms was established through a bag of words model, generating a features vector for each project site corpus $\{\mathbf{v}_{d_1},...,\mathbf{v}_{d_n}\}$, where each feature represents a word weighted by a TF-IDF weighting scheme [41]. Then, the mean of all features vectors sorted by the highest weighted features was calculated (Equation 1). Table 4 present the top 10 features retrieved $\{\mathbf{v}_{m_1},...,\mathbf{v}_{m_{10}}\}$. That is, the top 10 most common terms in the text of project URLs after removing irrelevant words such as stopwords. An example of a query derived using this heuristic is: "IMPACT Impact Measurement and Performance Analysis of CSR **project** -site:cordis.europa.eu -site:ec.europa.eu".

6 Heuristics tuning and performance

Each heuristic performance was measured and compared through *recall* (2), *precision* (3) and *f-measure* (4) metrics to evaluate the success of the proposed heuristic on identifying the project URLs of the test collection. The scores were measured by analyzing the Top 1, Top 3, Top 5 and Top 10 results obtained through the Bing Web Search API.

$$recall = \frac{|\{\text{relevant documents}\} \bigcap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \tag{2}$$

$$precision = \frac{|\{\text{relevant documents}\} \bigcap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \tag{3}$$

$$f\text{-}measure = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

Before comparing the performance between the described heuristics, the selection of common terms added to the heuristic **+Acronym +Title -Cordis -EC +CommonTerms** was tuned looking for the combination with the highest potential to provide the best performance. The following term combinations were tested: $\{project\}$,

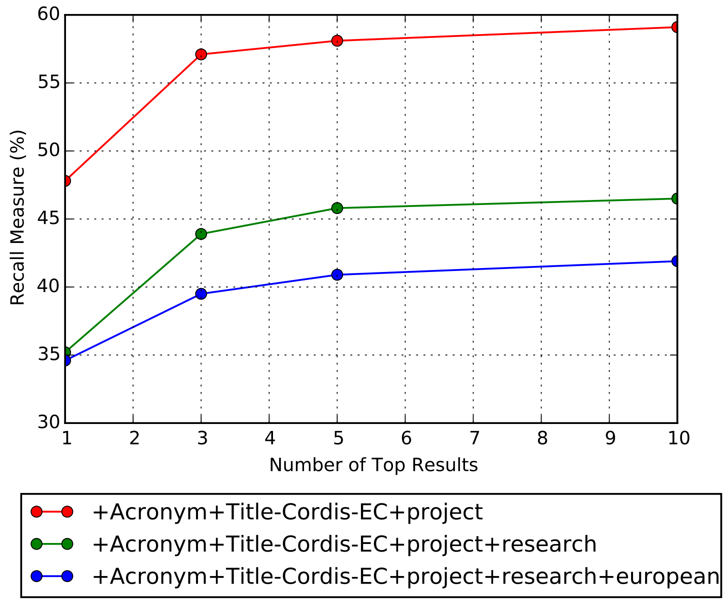

Fig. 2. Recall of heuristics using additional common terms in queries.

{*project*,*research*}, {*project*,*research*,*european*}. Based on the results presented on Figure 2, it was determined that the usage of only one term {*project*} provided the best results. Increasing the number of terms restricts too much the query scope obtaining lower recall values. Therefore, we decided to adopt only the addtitional common term *project* for the heuristic **+Acronym +Title -Cordis -EC +CommonTerms** and named it **+Acronym +Title -Cordis -EC +project**.

Table 5. Recall of each heuristic when identifying project URLs on the live-web.

	Top 1	Top 3	Top 5	Top 10
+Acronym +Title	44.0%	56.3%	*64.0%*	*66.0%*
+Acronym +Title -Cordis	44.9%	55.1%	58.1%	60.1%
+Acronym +Title -Cordis -EC	46.8%	56.1%	58.5%	60.5%
+Acronym +Title -Cordis -EC +project	*47.8%*	*57.1%*	58.1%	59.1%

Table 5 presents the score results for *recall* obtained for all the heuristics. As expected, it shows that increasing the number of top results retrieved increases the *recall* score. The heuristic with best recall (47.8%) at the TOP 1 results is the **+Acronym +Title +project -Cordis -EC**, but this is the worst heuristic at the Top 10 results (59.1% against **+Acronym +Title** 66%). Since this heuristic query contains more terms, it is more specific, becoming more precise at the Top 1 results, but the lack of generalization makes it worse with more results returned. Therefore, we conclude that is the most suitable heuristic when we aim to achieve more precise identification and retrieval of project URLs. The **+Acronym +Title** heuristic is the more general query and so it returns more results. It is most suitable when the objective is to obtain the highest coverage of project URLs to be preserved without limiting resources and preserving also some less relevant sites.

Table 6. Precision of each heuristic.

	Top 1	Top 3	Top 5	Top 10
+Acronym +Title	44.0%	*18.8%*	*12.8%*	6.6%
+Acronym +Title -Cordis	44.9%	18.4%	11.6%	6.0%
+Acronym +Title -Cordis -EC	46.8%	18.7%	11.7%	6.0%
+Acronym +Title -Cordis -EC +project	*47.8%*	*19.0%*	11.6%	5.9%

Table 6 indicates the *precision* scores obtained. As expected, they decrease as more results returned are considered because each query has only 1 valid result identified on the test collection. The heuristic that presented higher precision values was **+Acronym +Title -Cordis -EC +project** with 47.8%.

Table 7. f-measure of each heuristic.

	Top 1	Top 3	Top 5	Top 10
+Acronym +Title	44.0%	28.2%	*21.3%*	*12.0%*
+Acronym +Title -Cordis	44.9%	27.6%	19.3%	10.9%
+Acronym +Title -Cordis -EC	46.8%	*28.1%*	19.5%	10.9%
+Acronym +Title -Cordis -EC +project	*47.8%*	28.5%	19.3%	10.7%

The *F-measure* metric provides a combination of the *recall* and *precision* values. Those results are presented on Table 7 and show that **+Acronym +Title +project -Cordis -EC** has the highest score with 47.8% at Top 1.

7 Preserving R&D Websites: Experimental Crawl

The research previously presented was performed with the final aim of enabling the automatic identification and preservation of R&D Websites. Therefore, we applied the obtained knowledge to perform our our sites crawl on R&D websites. We decided to begin with a selective craw on FP7 Websites.

The experiments previously described were also tested using Google Custom Search Engine. This provided better results with an overall *recall* gain of 5% against Bing, but the limitation of 100 queries per day made it impracticable because the testing procedure of the heuristics was too slow. We believe that the ability to perform 5 000 queries/month of Bing Web Search API compensate for the slightly worse performance. For that reason Bing was the search engine that we used for the identification of new project URLs for R&D websites. The obtained results showed that the heuristic **+Acronym +Title -Cordis -EC +CommonTerms** achieved the best performance recovering project R&D URLs using the first result (Top 1), so it was the elected heuristic to apply to the incomplete FP7 projects dataset that presented 23 588 missing project URLs. The following work flow was executed to identify and preserve R&D project websites using the heuristics developed:

1. Extracted all project entries where *project URL* field was missing from the FP7 dataset;
2. Executed the heuristic **+Acronym +Title -Cordis -EC +CommonTerms** on FP7 projects dataset to recover missing URLs;
3. Used the newly identified URLs has seeds to the Heritrix crawler [38];
4. Harvested these project URLs and preserved this information.

After applying this workflow to the FP7 dataset, we identified 20 429 new URLs from the 23 588 entries with missing project URLs. That is, we improved the completeness of the CORDIS dataset by 86.6% regarding the project URLs information. About 3 159 entries did not return any URL, most probably because the project site does not exist any more, or never did.

Table 8. Data related to R&D project websites collected by the crawler for preservation.

Nr. project URL seeds	20 429
Nr. web files crawled	10 449 947
Nr. hosts crawled	72 077
Stored content size (compressed)	1.4 TB

These 20 429 new found project URLs were used as seeds to a new selective crawl that resulted on the collection of 10 449 947 web files, fulfilling a total of 1.4 TB of information compressed on ARC files as presented on Table 8. This selective crawl was configured to crawl all mime types, following links until 5 hops from the project URL seed, with a limitation of 10 000 files per site.

Figure 3 depicts the project URLs domain distribution on the crawl. Most of the crawled R&D project sites were hosted under the .EU domain. So, we measured the overlap between the preserved content using the **+Acronym +Title -Cordis -EC +CommonTerms** heuristic and the crawled content obtained from our previous .EU domain crawl [25]. Using the OpenSearch [16] API available at `arquivo.pt/resawdev`, we queried if the projects URLs obtained had been previously harvested. Only 9% of the retrieved R&D projects websites were previously preserved by the .EU crawl.

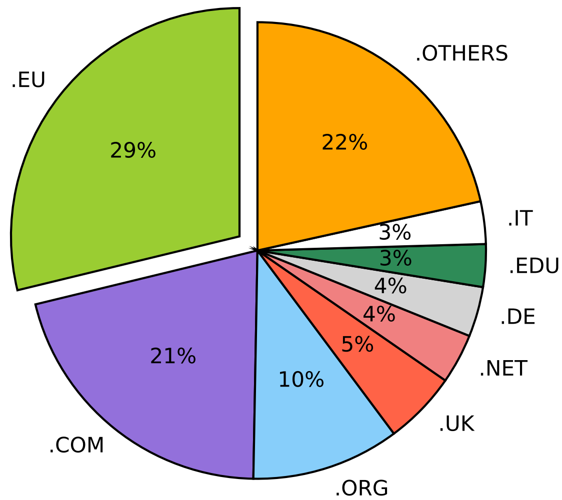

Fig. 3. Retrieved R&D projects websites domain distribution.

8 Preserving FP4, FP5 and FP6 R&D websites

After the experimental crawl of FP7 websites we performed another selective crawl to collect the FP4, FP5 and FP6 project websites. We applied the same heuristics to the FP4, FP5 and FP6 CORDIS Database and identified 37 707 new URLs that were used as seeds to the new crawl. The total number of documents preserved and storage used is presented on Table 9.

Table 9. Amount of data crawled from FP4, FP5 and FP6 websites.

Nr. Seeds	37 707
Nr. of web files preserved	40 674 738
Nr. of hosts crawled	1 035 001
Stored content size (compressed)	5.4 TB

This collection ran in two phases, the first phase started at 13 May 2016 and ended at 01 June 2016, with a duration of 19 days. The crawl target was the seeds discovered for projects that didn't have any project URL. The second crawl started at 14 July 2016 and it ran for 1 day, finishing at 15 July 2016. This crawl used as seeds the project URLs that still have valid content.

9 Preserving FCT funded project

A similar approach was applied to identify and preserve R&D project sites funded by the Portuguese funding agency FCT. The data source with information about the

Table 10. Wrongly filled project entry.

Acronym	Title
UI 408 - 2014	Incentivo - UI 408 - 2014

projects funded by FCT have a total of 11 996 project entries and none had information about the project URL. Also, the FCT data source has incoherences regarding the information that it made available. For instances, the acronym field almost never is filled. Also, some entries do not have any information regarding the project, with the acronym and title fields having not useful information as displayed on Table 10.

The heuristic used was the **+Acronym +Title**, but the data source rarely provided information regarding the Acronym for each project.

The described heuristic was applied and a total of 11 214 project URLs or related URLs was retrieved. After removing URLs duplicates, a total of 7 956 URLs was obtained and used as URL seeds.

The returned URLs were diverse. Manually inspecting the results we identified URLs pointing to national scientific repositories, research groups web pages, investigators personal pages and blogs that were involved in the projects.

A crawl was performed with this seed list. It started at 17 August 2016 and ran for 10 hours. The crawl was configured to follow 5 hops and download 1 000 files per site.

Table 11. Amount of data crawled.

Nr. Seeds	7 956
Nr. of web files preserved	600 721
Nr. of hosts crawled	11 546
Stored content size (compressed)	72 GB

Table 11 presents the total number of documents preserved and storage used by the crawl, a total of 600 721 web files was preserved with 72 GB of storage used.

10 Automatic filtering of project websites

The heuristics studied on Section 7 were conceived and tuned with the goal of maximizing the *precision* and recover missing project URLs from search engines. After the application of the best heuristic, it retrieved 20 429 FP7 project websites. From this list of results its unknown which were the really FP7 websites and which were not. We know that the heuristic is able to recover a FP7 project websites 49% of the times.

Verifying manually the retrieved URLs to filter those that are really project URLs requires too many human resources.

To overcome this situation an experiment to automatically verify if the retrieved URL is really a FP7 URL project or not has been tested.

The experiment tries to solve and minimize this problem using a machine learning binary classification technique to identify which websites are from FP7 work programme and which websites are not. The expected output from this experiment is a more fine grained list of FP7 projects websites and an evaluation of the methodology proposed.

10.1 Building the train and test dataset

A dataset is needed to train and test the classification algorithm. This is a binary classification problem, so the dataset was composed by 1 301 samples of FP7 projects websites and 425 not-FP7 projects websites. The classes group is defined as $\Omega = \{w_s, w_n\}$ where w_s stands for the FP7 projects websites class and w_n as the not-FP7 projects websites. Samples used to build the dataset have been taken from the following sources:

- Test collection build on Section 4 with FP7 project websites;
- No-FP7 projects websites that were returned by TOP1 **+Acronym +Title** heuristic against full test collection.

Table 12. Dataset build from samples extracted from CORDIS FP7 dataset and false positives returned from the heuristic.

	FP7 (w_s)	Not-FP7 (w_n)
Nr. of samples to test	301	125
Nr. of samples to train	1000	300
Total nr. of samples	1301	425

To train and test the algorithm, the dataset was split in a train dataset with 1 000 FP7 samples and 300 not-FP7 sample. The test dataset composed by 301 FP7 samples and 125 not-FP7 samples as described on Table 12.

10.2 Classification algorithm

The algorithm used was a multi-variate Bernoulli Naive Bayes [19,37]. The choice of this algorithm was justified by the following reasons:

- Its a well-known classifier with good performance on short text classification, the target is FP7 projects websites which don't have long text (homepages).
- Easy to train algorithm, with almost no parameter tuning.
- Has good performance with smaller number of features, our number of samples is low, so there is not enough data to use too many features.

$$\hat{w} = \arg\max_{w \in \Omega} P(w) \prod_{i=1}^{n} P(x_i \mid w) \tag{5}$$

$$P(x_i \mid w) = P(i \mid w) x_i + (1 - P(i \mid w))(1 - x_i), \qquad w \in \Omega \tag{6}$$

The general Naive Bayes algorithm classify a document based on the Equation 5, where x_i stands for the term i on the website x. Each Naive Bayes variant differs by the assumed the distribution model of $P(x_i \mid w)$. The multivariate Bernoulli distribution equation is shown at Equation 6.

For each home page text on the train dataset, a feature vector was computed and each feature represents a word frequency, English stop words were removed. A total of 1 300 word frequency vectors were generated, and the extracted vocabulary size reached a max dimension of 42 864 words with the train dataset. The computed vectors then are used to train the Bernoulli Naive Bayes algorithm, the algorithm needs a binary input so the features vectors were converted to binary to be used by the classifier.

10.3 Feature selection and performance measure

The algorithm does not need almost any parameter tuning, but the number of features need to be optimally selected to maximize its performance. To tune the number of features that should be used to train the classifier maximizing its performance, the following method was applied:

1. Increase incrementally by 10 the number of max top features used.
2. Measure the *f-measure*, *precision* and *recall* on the test dataset for each increment.

With each incremental step, the classifier performance is measured against the test collection. On a Machine Learning context the *precision*,*recall* and *f-measure* is defined below, with TP, TN, FP and FN standing respectively for True Positives, True Negatives, False Positives and False Negatives.

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{7}$$

$$precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

$$recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

$$f\text{-}measure = 2 \times \frac{precision * recall}{precision + recall} \tag{10}$$

The Figure 4 was generated with the described process. It shows the evolution of the classifier metrics when we vary the number of features used. We choose to use a max top features number of 800, with that number of features we maximize the *f-measure* score. More robust methods to select the best features can be tested, but as has been reported before, the objective was to gather quick insights.

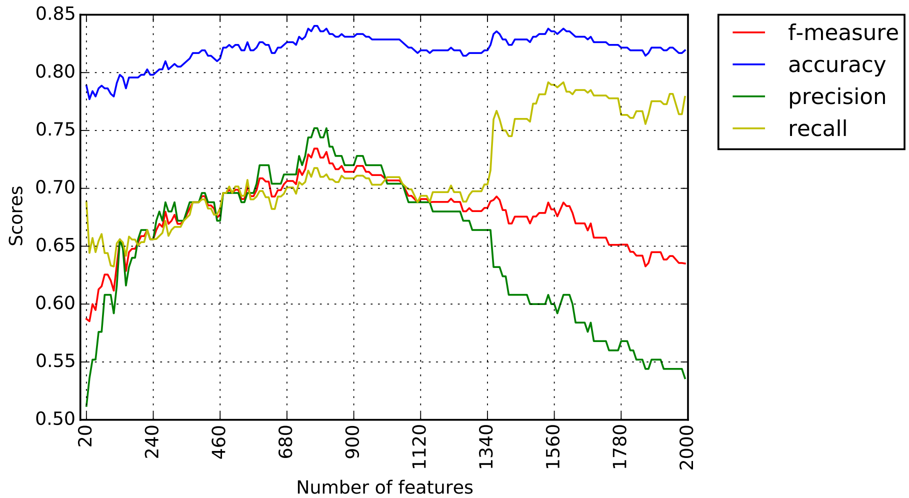

Fig. 4. Performance measure varying the number of features used.

10.4 FP7 projects classification results

The trained classifier was ran against the 22 429 retrieved results on the Section 7, classifying as FP7 projects a total of 5 613 entries with a 83% classification accuracy [11].

10.5 Improvements

The approach implemented was a fast and practical approach, the goal was not get the best results possible but to demonstrate some results and give some insight about the method of filtering of FP7 project websites with a machine learning classifier trained with the test collection.

Some improvements should be done to try get better results, and better evaluate the algorithms:

- Better feature selecting using information measures like entropy or mutual gain.
- Getting more no-FP7 projects samples to balance classes.
- Test more exhaustively different algorithms, SVM is a good candidate with better feature selection.
- Usage of cross validation to evaluate the results.

11 Conclusions

The objective of this work was to study and develop an automatic mechanism that would enable the identification of R&D project URLs to be preserved without requiring strong human intervention neither demanding computer resources, so that it could be replicated and applied by other organizations concerned with the preservation of scientific outputs such as libraries, archives or funding agencies. We analysed several heuristics that aimed to automatically identify missing project URLs using the search engines with information publicly available about the projects. The **+Acronym +Title +project -Cordis -EC** heuristic have the best performance using Top 1 results, being able to retrieve 47.8% of the missing projects URLs on the test collection. The *recall* score is improved with the usage of more retrieved results, reaching a score of 66% with the **+Acronym +Title** heuristic using the Top 10 results, but have the drawback of including more irrelevant results.

The **+Acronym +Title +project -Cordis -EC** heuristic was applied to try identify and recover the 23 588 project URL missing on the CORDIS data set from the FP7 programme. It has successfully retrieved 20 429 URLs with high potential of being the original project URL or related content. The newly identified project URLs were used as starting point to a preservation crawl using the Heritrix archiving crawler. Being able to crawl and preserve 10 449 947 Web files fulfilling a total of 1.4 *Terabytes* of information compressed on ARC files.

A selective approach using this method was also applied to FP4, FP5 and FP6 projects websites, preserving a total of 39 913 500 web files with a total storage size of 5.4 *Terabytes* on ARC files.

The same approach was applied to identify and preserve R&D project sites funded by the Portuguese funding agency FCT. The dataset available to conduct the heuristics was poorer regarding the available metadata. Yet the approach was able to identify important project related Web Sites like national scientific repositories, research groups web pages, investigators personal pages and blogs that were involved in the projects.

A crawl was performed with the identified Web Sites. It started at 17 August 2016 and ran for 10 hours. Preserving 600 721 web files and a total of 72 GB of information stored.

With the test collection built while studying the automatic identification heuristics, an exploratory experiment using Machine Learning classification algorithms to filter the URLs returned by this heuristics was conducted. An total of 5 613 entries were filtered with a 83 % classification accuracy. This methodologies to filter the URLs can be used when the resources available are more scarce, using a more selective range of URLs to crawl and preserve.

These R&D project websites content may have changed during their past lifetime, and this information is irrecoverable unless Web Archives holds past versions of these sites.

As societies evolve and become more aware of the importance of preserving born-digital content, it is expectable that R&D project websites will become systematically identified, archived and preserved during administrative work flows, so that the presented heuristics will become necessary only for exceptional situations. Meanwhile, automatic heuristics are crucial to preserve online scientific outputs.

12 Future Work

In future work these heuristics could be improved to reach higher levels of *recall*. One way to try improve this heuristics is to exclude more research network and funding websites that were not previously identified, for instance http://erc.europa.eu. Configuring the query to not return PDF format files could also improve the overall *recall* and contribute with more quality project URLs seeds, because PDF documents are not the project website and are not very rich sources to link discovery by harvesting crawls. Other methodologies and term combinations to extract describing words and improve query results also could be studied.

The test collection could be extended with additional results such as several relevant project URLs per each project entry. This extensions would accommodate situations such as projects that had several URLs across time or that provide several versions of the project URL in different languages.

References

1. Archive-It - Web Archiving Services for Libraries and Archives. https://archive-it.org/.
2. Arquivo.pt: pesquisa sobre o passado. http://arquivo.pt/.
3. Bing Search API Web — Microsoft Azure Marketplace. http://datamarket.azure.com/dataset/bing/searchweb.
4. CORDIS - EU research projects under FP4 (1994-1998) Datasets. https://open-data.europa.eu/en/data/dataset/cordisfp4projects.
5. CORDIS - EU research projects under FP5 (1998-2002) Datasets. https://open-data.europa.eu/en/data/dataset/cordisfp5projects.
6. CORDIS - EU research projects under FP6 (2002-2006) -Datasets. https://open-data.europa.eu/en/data/dataset/cordisfp6projects.
7. CORDIS - EU research projects under FP7 (2007-2013) Datasets. http://open-data.europa.eu/en/data/dataset/cordisfp7projects.
8. European Union Open Data Portal. http://open-data.europa.eu/en/data/.
9. FAROO - Free Search API. http://www.faroo.com/hp/api/api.html.
10. FCT - Fundação para a Ciência e a Tecnologia. http://www.fct.pt/index.phtml.en.
11. FP7 classified URLs list. https://github.com/arquivo/Research-Websites-Preservation/blob/master/classifier/filtered_projects_classification.csv.
12. Google Custom Search Engine. https://cse.google.com/.
13. International School of Information Science (ISIS). http://www.bibalex.org/isis/frontend/home/home.aspx.
14. Internet Archive: Digital Library of Free Books, Movies, Music & Wayback Machine. https://archive.org/index.php.
15. Library of Congress. https://www.loc.gov/.
16. OpenSearch. http://www.opensearch.org/Home.
17. Research resources and outputs. https://github.com/arquivo/Research-Websites-Preservation.
18. Test collection 300 samples. https://github.com/arquivo/Research-Websites-Preservation/blob/master/datasets/fp7-golden-dataset-300.csv.

19. The Bernoulli model. http://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html.
20. Time Travel. http://timetravel.mementoweb.org/.
21. Websites Archives of EU Institutions. http://www.eui.eu/Research/HistoricalArchivesOfEU/WebsitesArchivesofEUInstitutions.aspx.
22. S. G. Ainsworth, A. AlSum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How Much of the Web Is Archived? pages 1–10, 2012.
23. S. Alam, M. L. Nelson, H. Van de Sompel, L. L. Balakireva, H. Shankar, and D. S. H. Rosenthal. Web Archive Profiling Through CDX Summarization. *Research and Advanced Technology for Digital Libraries*, 9316:3–14, 2015.
24. ARCOMEM. Arcomem. https://web.archive.org/web/20130426060455/http://www.arcomem.eu/, October 2011.
25. D. Bicho and D. Gomes. A first attempt to archive the .EU domain Technical report. http://arquivo.pt/crawlreport/Crawling_Domain_EU.pdf, 2015.
26. British Library. UK Web Archive. http://www.webarchive.org.uk/ukwa/, 2011.
27. S. Chakrabarti, M. Van Den Berg, and B. Dom. Focused crawling: A new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11):1623–1640, 1999.
28. H. V. de Sompel, M. Nelson, and R. Sanderson. Http framework for time-based access to resource states – memento. RFC 7089, RFC Editor, December 2013.
29. D. Eastlake and P. Jones. Us secure hash algorithm 1 (sha1). RFC 3174, RFC Editor, September 2001. http://www.rfc-editor.org/rfc/rfc3174.txt.
30. E. Evangelou, T. A. Trikalinos, and J. P. Ioannidis. Unavailability of online supplementary scientific information from articles published in major journals. *The FASEB Journal*, 19(14):1943–1944, 2005.
31. M. Faheem and P. Senellart. Intelligent and adaptive crawling of web applications for web archiving. In *Web Engineering*, pages 306–322. Springer, 2013.
32. D. H.-L. Goh and P. K. Ng. Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology*, 58(1):15–24, 2007.
33. D. Gomes and M. J. Silva. Modelling information persistence on the web. In *ICWE '06: Proceedings of the 6th international conference on Web engineering*, pages 193–200, New York, NY, USA, 2006. ACM Press.
34. Internet Memory Foundation. Internet Memory Foundation. http://internetmemory.org/en/.
35. M. Klein and M. L. Nelson. Evaluating methods to rediscover missing web pages from the web infrastructure. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 59–68. ACM, 2010.
36. J. Kornblum. Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation*, 3(SUPPL.):91–97, 2006.
37. A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48, 1998.
38. G. Mohr, M. Stack, I. Ranitovic, D. Avery, and M. Kimpton. An introduction to Heritrix: An open source archvial quality Web crawler. In *4th International Web Archiving Workshop*, number 2004, 2004.
39. National and University Library of Iceland. Vefsafn - English. http://vefsafn.is/index.php?page=english, 2011.
40. A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, pages 1–12. ACM Press, 2004.
41. J. Ramos, J. Eden, and R. Edu. Using TF-IDF to Determine Word Relevance in Document Queries. *Processing*, 2003.

42. T. Risse, S. Dietze, W. Peters, K. Doka, Y. Stavrakas, and P. Senellart. Exploiting the social and semantic web for guided web archiving. In *Theory and Practice of Digital Libraries*, pages 426–432. Springer, 2012.
43. R. L. Rivest. The md5 message-digest algorithm. RFC 1321, RFC Editor, April 1992. http://www.rfc-editor.org/rfc/rfc1321.txt.
44. B. Sampath Kumar and K. Manoj Kumar. Decay and half-life period of online citations cited in open access journals. *The International Information & Library Review*, 44(4):202–211, 2012.
45. J. L. Shipman, M. Klein, and M. L. Nelson. Using web page titles to rediscover lost web pages. *arXiv preprint arXiv:1002.2439*, 2010.
46. A. S. Sife and R. Bernard. Persistence and decay of web citations used in theses and dissertations available at the sokoine national agricultural library, tanzania. *International Journal of Education and Development using Information and Communication Technology*, 9(2):85, 2013.
47. D. Spinellis. The decay and failures of web references. *Communications of the ACM*, 46(1):71–77, 2003.
48. O. Tajeddini, A. Azimi, A. Sadatmoosavi, and H. Sharif-Moghaddam. Death of web citations: a serious alarm for authors. *Malaysian Journal of Library & Information Science*, 16(3):17–29, 2011.
49. The National Archives. UK Government Web Archive — The National Archives. http://www.nationalarchives.gov.uk/webarchive/, 2011.
50. The Web Archive of Catalonia. The Web Archive of Catalonia. http://www.padi.cat/en.