Procedia Computer Science

Volume 80, 2016, Pages 1865–1874

ICCS 2016. The International Conference on Computational Science

Comparing electoral campaigns by analysing online data

Javier A. Espinosa-Oviedo[*†] Genoveva Vargas-Solar[†§**]
Vassil Alexandrov[*‡◊] Géraldine Castel[&]
BSC, Barcelona Supercomputing Centre [*]
CNRS, French Council of Scientific Research [**]
ICREA, Catalan Institution for Research and Advanced Studies [‡]
ITESM, Tecnológico de Monterrey [◊]
LAFMIA, French-Mexican Laboratory of Informatics and Automatic Control [†]
LIG, Laboratory of Informatics of Grenoble [§]
Université Stendhal, Grenoble 3 [&]
{espinosa, gvargas}@imag.fr, vassil.alexandrov@bsc.es, geraldine.castel@u-grenoble3.fr

Abstract
Our work addresses the influence of ICT technologies for campaigning purposes on the evolving dynamics of information flows from the eminently social beings that candidates are. Our approach combines an analysis of contents to technological and methodological concerns. In particular, this paper presents results concerning three of the data collections life cycle phases: collection, cleaning, and storage. The result is a data collection ready to be analysed for different purposes. The paper also describes our experimental validation for comparing political campaigns behaviour in France and the United Kingdom during the European elections in 2014.
Keywords: Big Data, Storage, Data analytics, Social Networks, Network science

1 Introduction

The use of information and communication technologies (ICT) in the political sphere is nowadays a key aspect for running electoral campaigns. In our work we focus on studied *candidate practices* in the context of the electoral campaigns, in particular the European elections in 2014 in the UK and in France. Our work rested on the premise that political parties indubitably belong to and interact with the information and communication society in which they evolve. Thus, our work addresses the influence of the growing adoption of ICT for campaigning purposes on the evolving dynamics of information flows from the eminently social beings that candidates are. Our approach combines an analysis of contents to technological and methodological concerns. The breadth of these objectives as well as the amount of data to be considered pointed to a need for collaboration between several researchers for sharing out tasks and bringing together expertise from various disciplines. From the point of view of British civilisation and political sciences, an interdisciplinary approach was therefore

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.480

a necessary scientific and technical prerequisite for being able to study practices over the course of several elections and combining data collected from heterogeneous sources. This challenge provided a contextualised framework guided by genuine application needs in which to develop solutions beyond the theoretical stage to Computer sciences and applied mathematics.

This paper presents results concerning three of the data collections life cycle: collection, cleaning and storing. The result is a data collection ready to be analysed for different purposes. In particular, in our experimental validation it has been used for comparing political campaigns behaviour in France and the UK during the European elections in 2014. Accordingly, the remainder of the paper is organized as follows. Section 2 describes related works concerning both projects that use ICT for analysing political data for decision making purposes (i.e., determining campaigning strategies) and for comparing also campaigning practices. Section 3 describes the overview of our approach including architectural hypothesis. Section 4 presents our collection/storage solution for providing continuous data processing and curating data collections so that they can go through different analytics processes. Section 5 shows preliminary analytics results and conclusions driven from the collected data analytics. Finally Section 6 concludes the paper and discusses future work.

2 Related work

The use of ICTs for political purposes is a relatively new research field as the first publications date from the end of the 1980s [1, 2], following experiments from a few isolated candidates in the United States, and then in a variety of other countries [3, 4]. Descriptive analyses of tools and practices were completed by reflexions on the impact of such an evolution on the democratic processes of those countries [5, 6]. Several campaigns were the object of scrutiny so as to reach a more detailed understanding of the topic [7–9]. The continuous evolution of tools has been paralleled by a shift in attention from sites to forums [10, 11], blogs [12, 13], or social networks [14, 15]. Most analyses have concentrated on top-down communication strategies initiated from party headquarters to local campaigning teams or on voters‟ reactions to party initiatives. Conversely, our work chose to concentrate on local practices at constituency level from individual candidates in the context of the 2014 elections to the European parliament both in France and in the UK.

The following lines compare our work with approaches concerning data collections design and preparation to support analytics processes. In the database domain, the problem of integrating databases from different sources is not new [16]. Heterogeneous data integration on relational systems where heterogeneity was related to both data structure and semantics [17] lead to important results addressing schema integration, query rewriting and optimisation [18]. In most of these proposals, data providers (heterogeneous or not) are known in advance or discovered [19] and integration is done assuming knowledge about the data structure [20], content, semantics [21] and constraints.

Then, the emergence of new kinds of data providers like services introduced new challenges particularly a matching problem [20]. The assumption was that a query represented a data integration requirement that could be fulfilled by one or several data services, not known in advance, that should be looked up in registries [22]. These kind of approaches assumed, for example, the fact that the description of the data and content provided by services was stored as meta data, the fact that services exported data in a pivot data model [23].

Data started to acquire “new” properties (more volume, velocity, variety) and with them emerged the need of building huge curated data collections out of data produced by different devices, under different conditions and that could be analysed [24, 25]. The challenge is to collect data continuously [26] and to ensure that collections can be used to perform analysis [27]: statistical, data mining, machine learning, deep learning and so on [28]. Works and tools include collection, cleaning, profiling and distributed storage [29]. Languages like, JaQL [30], Pig Latin [31], data cleansing and data mining techniques have been applied for this purpose. The objective is to complete data, to detect errors,

ensure freshness and also to have views of its content (e.g., data types and value distribution, and possible correlations and dependencies among attributes [32]). Resulting data collections are then stored according to different "sharding" techniques and sometimes they are correlated with other collections [33]. Data scientists can then decide which analytics technics [34] can be applied to extract information, infer models and knowledge from data collections. The challenge is to perform a continuous process as new data are harvested and as new insights are obtained about analysed data. In general, data processing is computationally expensive and it requires storage and memory resources, since algorithms are greedy and require data in memory. So, works have emerged trying to study how to deploy solutions in architectures and environments that provide such resources. A great deal of research and technology has been devoted to parallel programming models, languages and environments deployed in architectures like the cloud, the grid or high performance computing centres [35]. Our work addresses the construction of data collections giving a comprehensive view of their content, for supporting the decision making of data analysts and scientists willing to apply the most appropriate techniques that can lead to generate information and then knowledge. We maintain (without complete materialization) these views and all the sequence of data transformations done for preparing raw data to maintain some provenance properties without necessarily generating more data than those strictly required to ensure experiment reproducibility.

3 Overview of the approach

Figure 1 shows the overview of our approach consisting of the following data processing steps: collection, cleaning and storing. This process is recurrently executed since new data is produced, views can change and the organization on persistence support too.

- *This first phase* proposes data collection strategies specialized according to the type of data provider and they are implemented by services. As shown in figure 1, data collection is done according to different modes (push, pull) and at different rates particularly when data are produced continuously. Some providers are Web pages and blogs that contain and upload information so we crawled the content using Web scrapping and crawling techniques and tools.
- *The second phase* corresponds to a data analytics process that identifies, for each attribute of a given data structure, the distribution of the values within the collected data. It also identifies missing values and infers some proposals based on computed values distributions (e.g., using extrapolation), as well as discovering possible relations among data attributes (e.g., equivalence, functional dependency, temporal or casual correlations). Inference of types, missing and null values and dependencies is done considering a certain level of uncertainty. Inferred values are tagged with an estimated precision probability based on the sample used for computing these estimations. The second phase generates views that provide an abstract representation of the data collection contents.
- *The third phase* concerns data curation and includes: cleansing, preparation, storage of data and views. Depending on the characteristics of the data, and the cleaning and preparation process, views can be materialized and stored together with raw data. In this phase we make decisions on the best ways of sharding data across different nodes in a cluster. These decisions consider the probability of data to be accessed and processed together based on their possible dependencies. Given the volume of the initial collections it can be costly to migrate data from one cluster node to another. So, we deal with uncertainty related to attribute dependencies that will determine the probability of being accessed together for performing some kind of analytics. The quality of the meta-data inferred for computing views is a key aspect that impacts the way data are organized in storage support. Data organization can ensure performance and reduction of memory and communication resources consumption during the data analytics processes.

Our work aims at guiding the gathering, storing and scientific exploitation of online data both in France and in the UK, according to confidentiality and respect for private life. We refer to rules raised in France by the CNIL (National Commission on Freedom and Informatics) and constraints pertaining to data ownership rights born out of the emergence of big data generated online. We focus on Service Level Agreement (SLA) guided integration of heterogeneous sources with special attention paid to legal, provenance-based and privacy rules enforcement [36]. Adapting the collection, cleaning and curation according to SLA contracts considering juridical aspects is an original aspect of our proposal. The following Sections give details on the proposed strategies for each of the phases 1 and 2 described above.

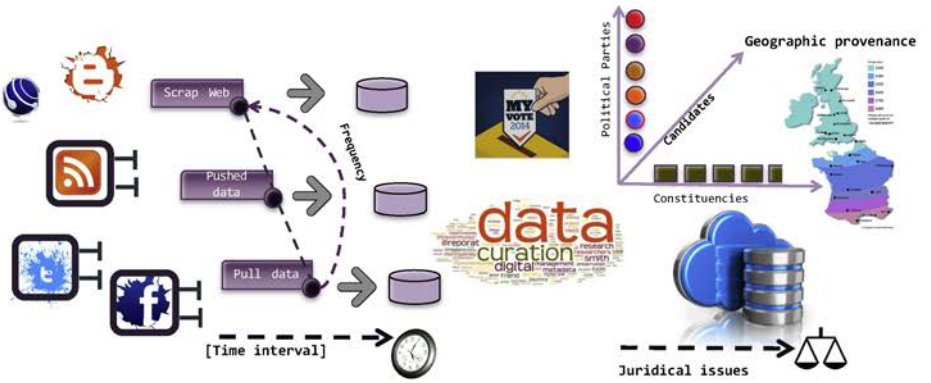

Figure 1 Data collection and curation overview

4 Collecting and analysing political data

We designed and built a database from the online production of a selection of candidates based on their blogs, sites, Twitter and Facebook accounts and interviews to confront theories derived from an analysis of raw online data with field-work data. Feedback from candidates would be collected from online questionnaires and semi-guided interviews applying novel crowdsourcing techniques combined with privacy constraints.[1] According to the type of data providers we used (i.e., Twitter, Facebook and official candidates web sites), we developed two general data collection strategies: *on demand* that uses data pulling techniques and *continuous* that supposes the production of data streams that are recurrently produced at some rate.

4.1 Collecting and archiving data

For collecting data, we assume that data providers are services implemented by a REST architecture or a SOAP API. These services are public and they can have specific SLA constraints, like the number of requests per hour (e.g., Twitter) or some authentication ones (e.g., Facebook). Other providers like pages or sites do not have explicit constraints but we assume that they are governed by privacy and authorship rights determined by their country of origin.

- *On demand* data providers have to be queried through a specific interface or crawled in order to harvest data. The frequency in which data is collected is specified in the program interacting with the provider. Figure 2 shows the general process implemented to interact with this type of providers. The consumer invokes the batch method through the network with its input data (a.1).

[1] The data collection referred in this paper does not include the information that was collected manually through interviews.

The speed of the network introduces the transfer time cost that is determined by the data size and the network's conditions (i.e., (g) latency and (h) throughput). Depending on the type of network, it can have a monetary price (e.g., 3G) also determined by data size. Once the method invocation arrives to the hosting device, the service provider receives the request and associates a predefined method invocation price (d). Afterwards, the method instance processes the request during an execution time (i.e., (b) method response time) which is determined by the method throughput (c) given by the amount of processed requests in a period of time (e.g., each minute) and the state of the device such as memory or CPU usage. The request implies the usage of the network interface, service provider and method execution. Those processes spend the battery of the device (i.e., battery consumption (f)) entailing a battery cost. Finally, the output (i.e., method response) is sent back to consumer through the same network and, as input data aforesaid, output data (a.2) contributes to data transfer time and to monetary cost. Both, input and output data define the data size measure (a).

- *Stream* providers work under a subscription strategy. A *continuous data provider* exports an method `subscribe()`used by a consumer to start receiving streams at some rate and for a given period of time (e.g., by executing an `unsubscribe()` method, for a predefined period of time, until something happens). Figure 2 shows the general process implemented to interact with this type of providers. The consumer invokes the continuous method through the network with its input data. Then, the method instance starts processing results and it sends the results every period of time (the so-called (j) production rate). The production rate can be determined by consumer needs. For instance, “give my current position every five minutes” where 'five minutes' is the expected production rate. Produced data is then sent to the consumer who processes it immediately or after a threshold defined by the number of tuples received, or the elapsed time, or a buffer capacity. This threshold is named (k) processing rate. Both production rate and processing rate impact the execution time cost, execution price cost, and battery consumption cost.

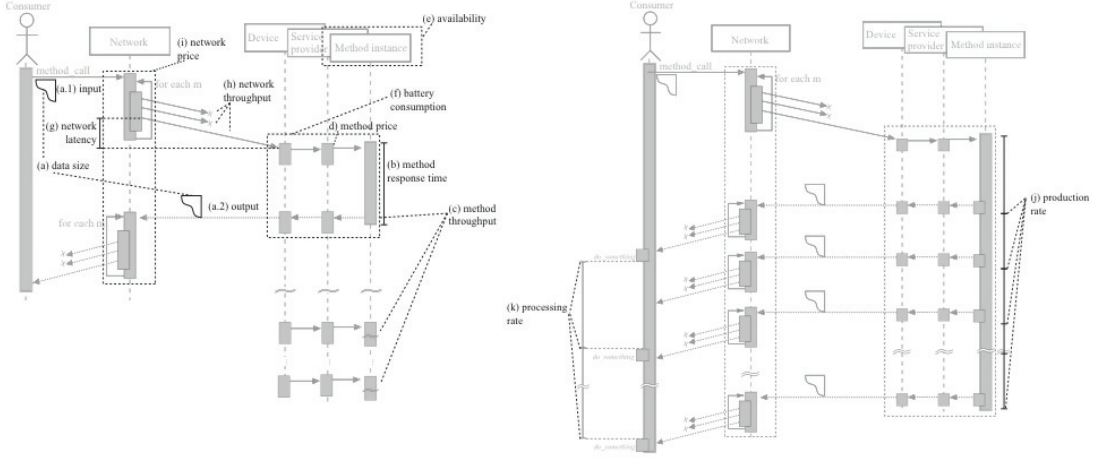

Figure 2 On demand and stream data collection from data providers

We assumed that providers are autonomous in the sense that they can modify their interfaces, authentication protocols and privacy and authorship rules whenever they want and our data collection services must deal with these changes. We do not have information about the production rate of the streams and changes in Web pages and sites. In a first approach we tuned the collection manually but we also collected information about services behaviour to automatize the tuning process and ensure to collect fresh non repeated data. We collected 30 Gigabyte of data about the European elections about candidates in the UK and in France. Data concern campaigns of 12 parties and 100 candidates, and

they concern only online activities reported in Twitter, Facebook and official sites, pages and blogs. We used JSON as data model and we then implemented document processing tasks to characterize the content of collected data.

4.2 Building and maintaining data views

Collected data compose raw data collections that must be analysed to get an abstract overview of the content in order to decide which cleaning and analytics techniques apply best to exploit them. The idea is not to transform data but to generate an abstract aggregated view and then eventually tag it with information that can be used for further data processing tasks that might generate transformed versions of these raw data. The view could be seen as a kind of schema in the relational world, but extracted *a posteriori* after having created a database.

In order to simplify we assume that data are represented as tuples and documents under a JSON like structure. Therefore, we define a view as a document that provides a description of every family of attributes of a raw document collection. For example, consider a collection of tweets from the European political campaign of candidates of the Labour party in the UK. A simplified version of these tweets has the following structure: `<"user", "date", "time", "location", "content">,` where almost all attributes are of type `String` and `"content"` can be of type `String, Image, Video, Sound`. Not all attributes are mandatory in every tweet and they can change of type from one tweet to another within the same collection.

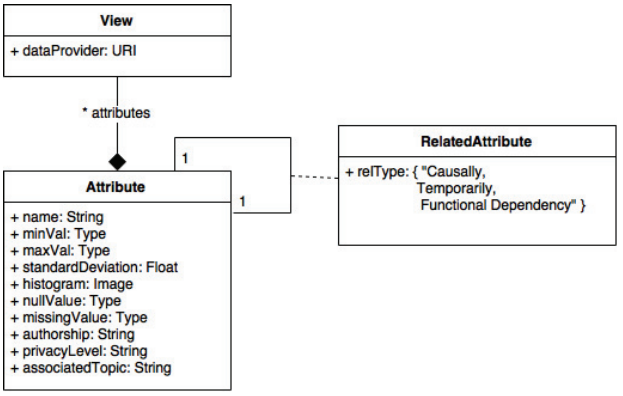

Figure 3 UML class diagram of the concept of View

As shown in the UML class diagram in figure 3, a `View` characterizes the content (document) provided by a given `dataProvider` as a set of attributes. For example, the view of our tweets collection consists of five classes of type `Attribute`, each class characterizing each of the attributes of a tweet, namely `<"user", "date", "time", "location", "content">.` The concept `Attribute` provides a snapshot of a given attribute's values domain for a given dataset. For example, the attribute `time` in our political data set ranged between the official initial date of campaigns for European elections to the date of the official announcement of results. An attribute within the dataset has maximum and minimum values, a standard deviation of the values assigned to the attribute in the different documents collected in the dataset, and the variation of values across the dataset elements represented by a histogram. Within a dataset an attribute can have null and missing values that must be inferred in order to characterize its domain type as precisely as possible. Indeed, many data collections represent missing values by dummy values and therefore we want to represent those cases. For example, in our political tweets collection, the attribute location was not always associated with a value. Since such values are inferred out by analysing a dataset such values have

associated precision probabilities that can measure uncertainty. In our political tweets, in some cases it was Twitter that associated an empty value.

A value of an attribute of a document in a dataset can have an author and it can be protected by an authorship licence, some privacy level and it can belong to a thematic classification. In the case of the tweets and particularly photographs they belong to Twitter and to the user. For images this was a huge assumption because some candidates use photographs that do not have recognized authors. We avoided this problem but this is part of open issues regarding juridical aspects studied in our work. Anyway, authorship and privacy level represent the condition in which the value of an attribute is produced and can be consumed. Concerning thematic classification, in the case of our political tweets we used hashtags within the contents to classify the tweets by topic. The content of a tweet was parsed looking for hashtags that could help to group the tweet and then try to compare topics with key words of the political proposals or the official speech of the corresponding party. We could also see that some trendy topics in Twitter were not that trendy in sites, and blogs. Again tweets classification is associated to uncertainty and was associated to probabilistic measures.

Finally, an attribute can be related to other attributes within a document with different relationship types: functional dependency, temporal and causal dependencies. For example, in the case of a tweet, the temporal attribute of the tweet has a temporal dependency with the temporal attribute of the replies or retweets. Replies and retweets should always be published later as the initial tweet otherwise there is an error. Relationships are computed using techniques stemming using numerical measures estimated from values. Some are more or less easy to identify for example we use hashtags to determine “semantic” similarity that we interpret as tweets using the same hashtags. Network science techniques provide a variety of strategies that can be used for providing a comprehensive view of the political campaigns done on Twitter and other social networks. Relationships that are deduced are tagged with probabilities and with measures that represent the influence of tweets towards others.

We associate a visual representation of a view that can be presented to a data scientist who can validate its contents. Once a data collection has been profiled with a view, the view can be stored, completed and modified and used for supporting a decision making process of a data scientist who will decide which analytics techniques to apply in order to extract knowledge and drive conclusions and models about a given subject.

5 Analysing and comparing political data

Based on our approach we built a system to analyse and compare campaigns in UK and France of the European elections in 2014. As shown in figure 4 recent advances in the field of social network analysis and data visualization could be put to use to chart the nature and direction of information flows from the original producer. Sharing of results with candidates in that respect could make access to them easier and help set up a mutually beneficial relationship resulting in both scientific advances and the eventual development of tools relevant to political activity both in France and in the UK and tailored to existing needs rather than imported from other sectors of activity such as the business sector which is frequently the case in matters of political communication.

Such tools could be based on social network analysis to maximize impact of online communication beyond “influence marketing‟ targeting allegedly influential individuals to voters. They could be contents aggregators or information watch instruments so as to enable a party and a candidate to centralise in real-time information from various sources pertaining to a specific topic or person. They could also facilitate multi-channel broadcasting of contents (see figure 5).

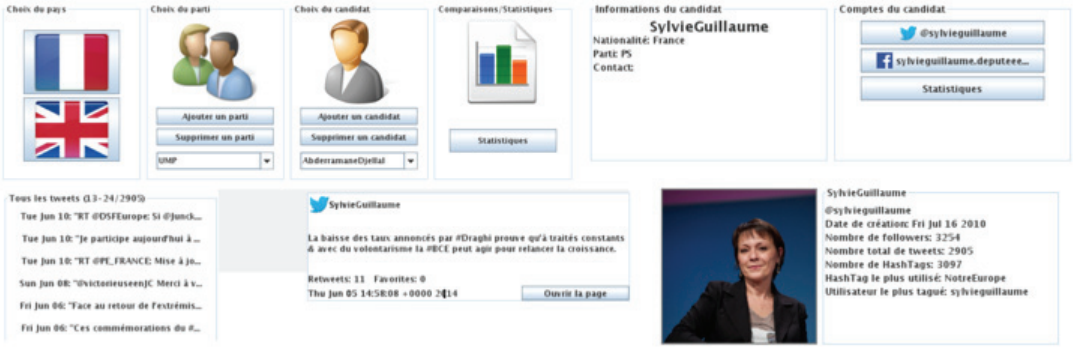

Figure 4 Profiling a candidate's campaign on social networks

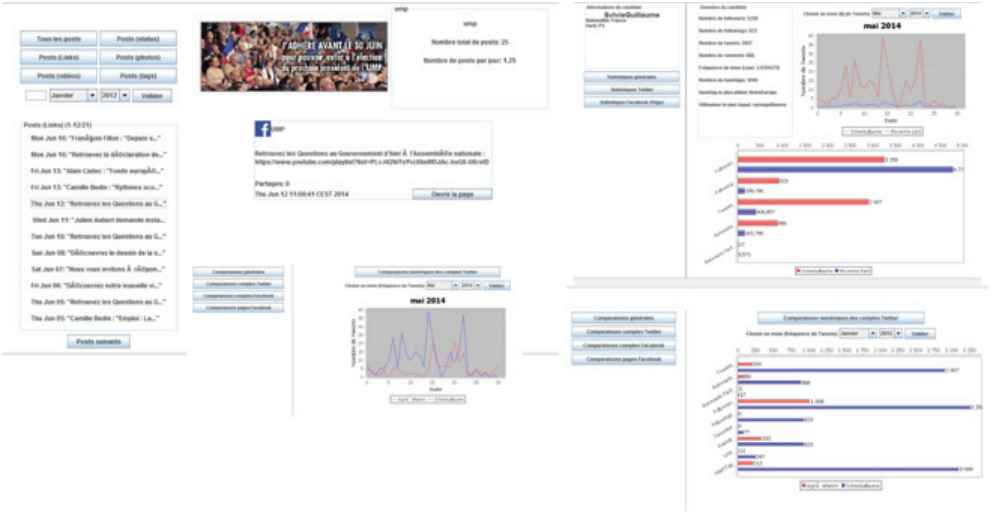

Figure 5 Statistical profile of politician's campaigns on social networks and Internet

6 Conclusions and future work

This paper introduced our approach for building and curating political data collections and preparing them for participating in analytics processes. We rather propose a workflow that includes activities addressing data collection, cleaning and curation based on the notion of views. These activities are guided by SLA criteria particularly juridical ones that control the way these processing phases are performed according to the type of data, the conditions in which they are produced and consumed. The activities consume computing, memory and storage resources at different scales depending on the volume of data and the complexity of the algorithms used. So, some of them are implemented under the map-reduce programming model and implemented and executed on cluster like architectures, using existing tools. Our first contribution in this paper regards the strategies used for characterizing and inferring data content through the notion of view combined with SLA constraints rather than the performance of these tasks. Some inference had to deal with uncertainty that we addressed associating accuracy probabilities to inferences so as to guide the data scientist in her further data analytics design.

The increasing use of ICTs by political candidates generates an unprecedented volume of technologically-mediated online information commonly referred to as “Big Data” which presents a

number of challenges. The necessity to manage and organise such large quantities of digital contents into meaningful items is among them and has become increasingly pressing for social scientists desirous to make use of such wealth of data but facing difficulties in matters related to gathering, storing and retrieving information on so large a scale. We aim to map out the diversity of uses of the internet for political campaigning in a variety of contexts and thus provide a comprehensive overview of practices and expectations from candidates in that respect so as to assist with the development of technologies based on practical and genuine needs in this area. Identifying potential applications and partners for this type of development shall be another objective and the 2017 French presidential election can offer an opportunity to help develop those tools.

References

1. Abramson, J.B., Arterton, F.C., Orren, G.R.: The Electronic Commonwealth: The Impact of New Media Technologies on Democratic Politics. Mich. Law Rev. 87, 1393 (1989).

2. Downing, J.D.H.: Computers for Political Change: PeaceNet and Public Data Access. J. Commun. 39, 154–162 (1989).

3. Auty, C., Nicholas, D.: British political parties and their web pages. Aslib Proc. 50, 283–296 (1998).

4. Hoff, J., Horrocks, I., Tops, P.: Introduction: New technology and the “crisis” of democracy. Democr. Gov. New Technol. Technol. Mediat. Innov. Polit. Wester Eur. 1–10 (2000).

5. Grossman, G.M., Krueger, A.B.: Economic Growth and the Environment. Q. J. Econ. 110, 353–377 (1995).

6. Margolis, M., Resnick, D.: Politics as Usual The Cyberspace `Revolution’. SAGE Publications, Inc (2000).

7. Gibson, R., Ward, S.: A Proposed Methodology for Studying the Function and Effectiveness of Party and Candidate Web Sites. Soc. Sci. Comput. Rev. 18, 301–319 (2000).

8. Foot, K.A., Schneider, S.M.: Web Campaigning. Polit. Psychol. 29, 463–466 (2008).

9. Vaccari, C.: From the air to the ground: The Internet in the 2004 US presidential campaign. New media Soc. 10, 647–665 (2008).

10. Wojcik, S.: Les forums électroniques municipaux,espaces de débat démocratique ? Sci. la Société Démocratie locale Internet. 107–125 (2003).

11. Marcoccia, M.: Les webforums des partis politiques français : quels modèles de discussion politique ? Mots. Les langages du Polit. [En ligne], URL http//mots.revues.org/512. 49–60 (2006).

12. Greffet, F.: Les blogues politiques. Enjeux et difficultés de recherche à partir de l’exemple français. Commun. Inf. médias théories Prat. 25, 200–211 (2007).

13. Gadras, S.: Public Sphere and Political Communication: how Does the Public Sphere Evolves with the Development of ICTs in French Local Politics? The European Public Sphere: From critical thinking to responsible action. Brussels, Peter Lang (2012).

14. Jackson, N., Lilleker, D.: Microblogging, Constituency Service and Impression Management: UK MPs and the Use of Twitter. J. Legis. Stud. 17, 86–105 (2011).

15. Margaretten, M., Gaber, I.: The Crisis in Public Communication and the Pursuit of Authenticity: An Analysis of the Twitter Feeds of Scottish MPs 2008-2010. Parliam. Aff. 67, 328–350 (2014).

16. Dong, X.L., Srivastava, D.: Big data integration. 2013 IEEE 29th International Conference on Data Engineering (ICDE). pp. 1245–1248. IEEE (2013).

17. Lara, R., Lausen, H., Arroyo, S., Bruijn, J. de, Fensel, D.: Semantic Web Services: description requirements and current technologies. International Workshop on Electronic Commerce, Agents, and Semantic Web Services (ICEC2003). , Pittsburgh, PA, USA (2003).

18. Halevy, A.Y.: Answering queries using views: A survey. VLDB J. 10, 270–294 (2001).

19. Dong, X.L., Berti-Equille, L., Hu, Y., Srivastava, D.: Global detection of complex copying relationships between sources. Proc. VLDB Endow. 3, 1358–1369 (2010).

20. Cuevas-Vicenttín, V., Zechinelli-Martini, J.L., Vargas-Solar, G.: Andromeda: Building e-Science Data Integration Tools. Proc. of the 17th Int. DEXA Conference. pp. 44–53. , Kraków, Poland (2006).

21. Osborne, F., Motta, E.: Klink-2: Integrating Multiple Web Sources to Generate Semantic Topic Networks. Proc. of the 14th Int. Semantic Web Conference (ISWC 2015). pp. 408–424. Springer International Publishing, Bethlehem Pennsylvania, USA (2015).

22. Meshkova, E., Riihijärvi, J., Petrova, M., Mähönen, P.: A survey on resource discovery mechanisms, peer-to-peer and service discovery frameworks. Comput. Networks. 52, 2097–2128 (2008).

23. Rekatsinas, T., Dong, X.L., Getoor, L., Srivastava, D.: Finding Quality in Quantity: The Challenge of Discovering Valuable Sources for Integration. Proc. of the 7th Biennial Conference on Innovative Data Systems Research (CIDR 15). , Asilomar, California, USA (2015).

24. Adiba, M., Castrejón, J.C., Espinosa-Oviedo, J.A., Vargas-Solar, G., Zechinelli-Martini, J.-L.: Big Data Management: Challenges, Approaches, Tools and their limitations. Networking for Big Data. CRC Press (2015).

25. Labrinidis, A., Jagadish, H. V.: Challenges and opportunities with big data. Proc. VLDB Endow. 5, 2032–2033 (2012).

26. Ma, M., Wang, P., Chu, C.-H.: Data Management for Internet of Things: Challenges, Approaches and Opportunities. 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing. pp. 1144–1151. IEEE (2013).

27. Vargas-Solar, G., Espinosa-Oviedo, J.A., Zechinelli-Martini, J.L.: Big continuous data: dealing with velocity by composing event streams. Big Data Concepts, Theories and Applications. Springer Verlag (2015).

28. Shah, I., Sheth, A.: INFOHARNESS: managing distributed, heterogeneous information. IEEE Internet Comput. 3, 18–28 (1999).

29. Barnaghi, P., Sheth, A., Henson, C.: From Data to Actionable Knowledge: Big Data Challenges in the Web of Things. IEEE Intell. Syst. 28, 6–11 (2013).

30. Beyer, K.S., Ercegovac, V., Gemulla, R., Eltabakh, M., Kanne, C.-C., Ozcan, F., Shekita, E.J.: Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. Proc. VLDB Endow. 4, (2011).

31. Olston, C., Reed, B., Srivastava, U., Kumar, R., Tomkins, A.: Pig latin: a not-so-foreign language for data processing. Proc. of the 2008 ACM SIGMOD Int. Conference on Management of data (SIGMOD'08). ACM Press, Vancouver, Canada (2008).

32. Park, J., Brenza, A.: Evaluation of semi-automatic metadata generation tools: A survey of the current state of the a. Inf. Technol. Libr. 34, 22–42 (2015).

33. Grolinger, K., Higashino, W.A., Tiwari, A., Capretz, M.A.: Data management in cloud environments: NoSQL and NewSQL data stores. J. Cloud Comput. Adv. Syst. Appl. 2, 22 (2013).

34. Cugola, G., Margara, A.: Processing flows of information. ACM Comput. Surv. 44, 1–62 (2012).

35. Di Stefano, M.: Distributed data management for Grid Computing. John Wiley & Sons (2005).

36. Bennani, N., Vargas-Solar, G., Ghedira, C., Souza-Neto, P., Carvalho, D.: Can Data Integration Quality be Enhanced on Multi-cloud using SLA? Proc. of the DEXA'15 Conference. , Valencia, Spain (2015).