An Application of Computational Collective Intelligence to Governance and Policy Modelling

Tien Van Do[1], Tamas Krejczinger[1], Michal Laclavik[4], Nguyen-Thinh Le[3], Marcin Maleszka[2], Giang Nguyen[4], and Ngoc Thanh Nguyen[2]

[1] Department of Networked Systems and Services, Budapest University of Technology and Economics, Hungary
[2] Institute of Informatics, Wroclaw University of Technology, Poland
[3] Department of Informatics, Clausthal University of Technology, Germany
[4] Institute of Informatics, Slovak Academy of Sciences, Slovakia

Abstract. The spread of social media provides a great opportunity to enhance the transparency, participation and collaboration in modern democracies. Since nothing is perfect, a best practice engineering approach can be used to continuously monitor processes and operations that are applied in increasing the participation of people and improving public service provision. This paper outlines some basic concept of our initiative called "Knowledge Management Tools for Quality of Experience Evaluation and Policy Modeling-KNOWN" that aims at the modeling and analysis of data collected from social media and other online sources. The purpose is to provide quantitative information and feedbacks regarding the quality of open government and public service provision.

1 Motivation and Main Goal

To organize and run a modern, democratic and well-developed society, huge costs are needed. The tax payers' (people and companies) money are used to organize public services, build infrastructure, stimulate economy, etc. Therefore, it is the joint and long-term interest of the society for a better future that more and more people are involved in a political process. However, people are less interested in common issues in new democratic and developed democratic countries as well, which are due to various reasons (e.g., the complex political processes, the lack of transparency in decision-making, the political inactivity, etc.). Furthermore, the quality and the efficiency of public service (to go to the local government to arrange some things –e.g., to obtain a new passport, identity card, etc.) provision are often questionable.

Therefore, processes, measures and best of practices are needed to increase the participation of people and to improve public service provision, which principles are laid by the Open Government Partnership (www.opengovpartnership.org) initiated in 2011. Since the establishment of the initiative, a number of countries have agreed to join forces in the Open Government Partnership to increase transparent and participatory government.

N.T. Nguyen, T. van Do, and H.A. Thi (Eds.): *ICCSAMA 2013*, SCI 479, pp. 237–249.
DOI: 10.1007/978-3-319-00293-4_18

The spread of social media (Facebook, MySpace, Twitter, Youtube, Flickr, Foursquare, Wiki and Google Doc) can provide a great opportunity to implement and to enhance the three principles (transparency, participation and collaboration) of open government [15]. Indeed, there are various examples that local governments and public organizations have a presence on Facebook, Twitter, and Youtube [6][7].

Furthermore, the research on monitoring open government based on data from social media is still in its infancy. The information extraction and the analysis of social media (that are built using Information and Communication Technology--ICT) can provide an efficient opportunity to quantitatively monitor and to evaluate public service provision and open government as a consequence of human activities and decisions, which provides a motivation for this work.

To efficiently support the efforts by the governments and agencies of the Open Government Partnership, we propose a joint effort and initiative called "Knowledge Management Tools for QoE Evaluation and Policy Modeling-KNOWN". The KNOWN project applies computational collective intelligence to model and analyze "open government-related" data collected from social media. The purpose of the modeling and analysis is to provide quantitative information and feedbacks regarding the quality of open government and public service provision. To our best knowledge, this proposal is a pioneering initiative in the field. To raise awareness to the problems, this paper presents a concept and proposed methods that will be applied in our initiative.

The rest of this paper is organized as follows. Section 2 covers the challenges of the initiative, describes the applied methodology. Section 3 presents the overview of the software tools. Finally, Section 4 provides general conclusions of the initiative.

2 Concept, Approach and Methodology

2.1 Challenges

We are convinced that intangible aspects related to human activities, society, ways people (politicians and citizens), society and economic environments interact with each other, should be taken into account (see Fig. 1.). This leads to the following major challenges:

- What are intangible aspects that are to be considered?
- What quantitative measures can be proposed to describe and to monitor the performance of public service provision?
- How computable models can be constructed that are able to characterize intangible aspects and processes?
- How can we determine the parameters of computable models?
- How accurate are predictions based on computable models (how can we validate models based on scientific evidence and to obtain scientific evidence)?

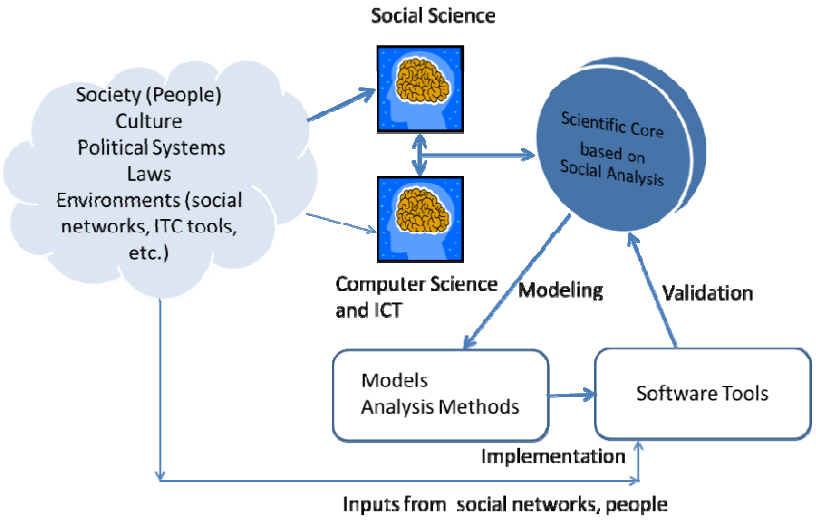

Fig. 1. The context and the environment, and the concept sketch of the KNOWN project

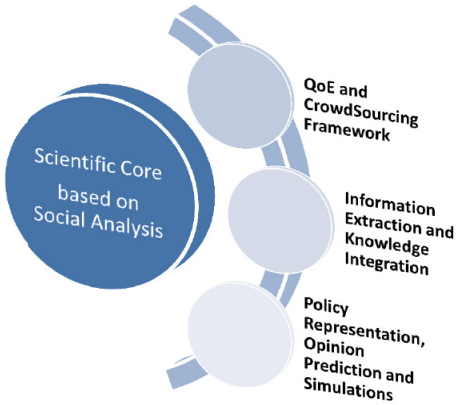

Fig. 2. The scientific methods

2.2 Concept

To achieve our goals, a systematic approach (Fig. 1) is performed as follows.

- Understanding the details of problems/challenges through application scenarios: To really serve the real needs we should understand problems in depth. We have to identify exactly scenarios and opportunities to identify the needs of society and end-users with the aim to improve the life of citizens, the participation in policy decisions, etc.
- Research directions are to be identified and ICT toolsets developed to solve problems that emerge in the scenarios. That is, the KNOWN project proposes and develops algorithms and methods that will be implemented in the

KNOWN software framework. Then, the methods and procedures developed will be subjected to extensive empirical validation. The test and validation of models and methods will be performed through comparing results obtained by mining and analysis of data collected from social media and results collected by the classical opinion polling. The comparison will give feedbacks to improve models and methods.

The scientific objectives to be achieved are categorized into three main frameworks as follows (Fig. 2):

- A Quality of Experience (QoE) framework allows us and the users of our software to define the mapping of opinions into quantitative measures
    - The survey and classification of a set of QoE (Quality of Experience) measures to evaluate policy decisions and public service provision,
    - The specification and investigation of methods to get qualitative solutions from crowd-sourcing techniques that are used in public service provision,
- a framework that is capable to perform large scale data extraction, processing and knowledge integration
    - The development of methods and procedures for the large scale extraction of information from various online sources,
    - The proposal of procedures to build semantic knowledge graph and integrate knowledge,
- a modelling, prediction and simulation framework that incorporates scientific techniques to evaluate and predict QoE measures and opinions based on extracted and processed data.
    - The definition of methodologies (argumentation graphs, conflict resolution) for representing policy and mining opinions, the use of Sentiment Analysis (SA) and Multidimension Scaling (MDS) for visualizing opinions and exploring similarities or dissimilarities in opinions,
    - The development of methods for evaluating and predicting the QoE measures and opinions based on mining data,
    - The development of stochastic models that are able to characterize the dynamic and stochastic behaviour aspects of participants (voters and politicians) in Governance and Policy based on mined information and data from various sources. We will validate proposed stochastic models and provide a tool to compute parameters using real data collected from social networks, crowd-sourcing and collaborative feedbacks in response to policy decisions.

Functions incorporating scientific methods and framework are proposed to be implemented in a scalable and extendable software architecture using advanced software engineering [13] and service oriented software architecture [1].

2.3 Scientific Methods

To achieve our aims, we plan to apply and to develop advanced computational collective intelligence methods. In what follows we outline some key methods and framework.

QoE framework
Lee and Kwak [7] have proposed an open government maturity model to access and guide open government initiatives after field studies regarding U.S. federal healthcare administration agencies. They closed their study by concluding:

"*Furthermore, open government metrics are currently under-developed. Future research needs to develop reliable and valid metrics to measure and demonstrate the return on investment in open government.*" [7]

Therefore, we will do research on defining a set of metrics we term as the collection of Quality of Experience (QoE) measures that can be used to characterize aspects and public opinions related to open government. Therefore, QoE measures/metrics can be interpreted as the quantitative measures of public opinions in the general sense. Furthermore, QoE measures should reflect the true feelings of people from their perspective when they watch policy processes, participate in some aspects of open government and use public services. We are aware this is the complex issue since mapping involves subjective judgements as well. Therefore, we will work out a QoE framework that is configurable by the users of the KNOWN software framework using the XML schema.

Information Extraction and Knowledge Integration
Methods suitable to be used for this purpose include: A/B testing, association rule learning, classification, cluster analysis, crowdsourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, predictive modelling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualisation.

Based on the current state of the art in the field of information retrieval [8], the Semantic Web [2], processing of information networks and graphs [10] and unstructured text [9], as well as our previous experience in these areas, we would like to develop new methods which can integrate, process, search and visualize structured and unstructured sources and deliver them as knowledge bases. Text is still the most used medium for information sharing and communication, available ubiquitously on the Web, emails or new social media as well in organizational digital assets. This textual data often points to graph/network data through Web links, communication links, transactions or social links and tags, but describes a large part of their knowledge in unstructured textual form as well. In the area of natural language processing and extraction of information, we shall develop methods for transforming text to structured semantic data, such as tags, annotations, semantic trees and graphs.

We have to model knowledge graphs from text and network data. This research will improve and generalize our existing approach applied successfully on business data such as emails and documents. We will model knowledge graphs from textual interconnected documents. The novel approach of exploitation view and semantic search will be also developed with the aim to improve the quality of collected data and gained structures.

The process of creating a knowledge graph or an ontology based on different data sources requires attention to problems known in the knowledge integration area. In particular, the following issues must be addressed:

- Eliminating data inconsistency between sources.
- Integratiing different sources into a single knowledge graph.
- Eliminating knowledge inconsistencies in output graph.

All issues may be addressed in a single integration algorithm, but it is preferable to approach them independently to improve the overall quality of the result.

Eliminating inconsistency between different data sources is a process well known in the integration theory area, but generally not considered for more complex structures, like graphs. An example of such inconsistency would be if the same user states pro-A in one source and against-A in another (note that at data level this conflict is direct; on knowledge level such conflict may be hidden and require additional inference to determine). For simple structures such as conjunctive or multi-value structures, we have to solve the inconsistency [11]. The solution may be a single statement of opinion, a statement of uncertainty or eliminating the issue from this user's representation (note that on the knowledge level other solutions are possible). This method may be used, among other things, to eliminate uncertain or fake users.

The second issue, integration of different sources into a single knowledge graph, presents another kind of challenges. The most important one is the size of the dataset under processing. In a related field of ontology matching, the challenge of large scale ontology matching is still an open research question. The largest previously processed graph structures, where mapping related elements between different sources was done, were cross-lingual cases in [3] consisting of tens of thousands of entities. In this project, large social networks may be considered, containing up to hundreds of thousands of users. As each user may possess multiple opinions for mining, the overall graph may contain number millions of entities.

The third issue is related to the first one, but operating on knowledge level requires another set of tools for eliminating inconsistencies. If some inconsistencies are not detected on data level, then more complex analysis on a knowledge level allows finding those conflicts. This step will be responsible for finding fake profiles, which are not consistent with others. An example of such inconsistency would be if the same user states A in some issue and states B in some related issue but these two opinions A and B are mutually exclusive (note that in the first step it was not possible to find this without analysis on the semantic level). Another problem appears when we get empty opinions during the integration process. Based on collected data we can infer missing opinions. Similarly like in the first step, we need multiple methods of solving inconsistencies in ontologies and resolving conflicts which may be used to eliminate fake users, empty opinions and other uncertainties.

Graph-Based Policy Representation and Policy Simulation

A policy can be considered as a principle or a rule which serves to achieve a rational impact. In order to simulate the impact of a policy, it is required to represent a policy. Since a policy is released on the basis of several supporting arguments. We need to support a policy maker (e.g., a politician) to sketch her argumentation process. One approach of representing the argumentation process is using an argumentation graph which may contain elements such as: facts, contra arguments, pro arguments, and hypotheses. There exist various computer-supported tools for argumentation [12]. The benefits of using a computer-supported argumentation tools include: 1) an incorrect argumentation process (e.g., a cyclic argumentation) will be detected automatically by the system, and 2) the graph-based representation of a policy can be used directly for simulation. We will investigate which tool is appropriate to represent policies and whether the tool can be integrated into the tools framework of this project.

After collecting data and transforming them into representations which capture semantics, we use these data for simulating reaction of people on a policy. The classification of opinions can be more complex. We will develop an algorithm to calculate the pro-/contra-scale for each individual opinion with respect to a policy. For this purpose, Bayesian networks can be deployed. The goal is to visualize how many people agree with the policy to which extent. Alternatively, methods [11] for solving inconsistencies in multi-attribute and multi-value data may be modified to achieve similar results. The advantage of this alternative method is that it groups the users without the additional processing step for calculating the scale. On the other hand, compared to the Bayesian network approach, it is significantly harder to influence the number of positions on the scale (if more or less is required).

According to the social impact theory, social influence is one of the pervasive forces that operate in groups and societies [5]. That is, each individual person can influence the opinion of her neighbors. This theory suggests that the amount of impact other people have on an individual depends on three parameters: 1) the number of people influencing or being influenced, 2) the strength of these people, and 3) their immediacy to each other. The strength of each individual represents the ability to influence other's opinions. One challenging problem is detecting the strength of each individual in a social network. For example, we can assume that a Facebook user who has more friends can be considered more influential (i.e., she has more strength) than a user with less friends. This issue deserves investigation in this project. The third parameter, the immediacy, is relevant in the social impact theory, because people interact most often and are mostly influenced by those who are close to them (such as family members, friends, and colleagues) and their neighbors (i.e., those who live close to them in physical space). The distance between the two individuals in social space can be mapped to their immediacy. Whether in virtual social networks from which we intend to collect data, the immediacy can also be determined in a similar way as in physical space, needs further investigation. Once we have determined values for these three parameters, the number of people in a social network, the strength of these people, and their immediacy, we are able to model and to simulate the change of opinions of people until their opinions reach a constant state. In order to model the opinions of people in a social network Stocker et al. [14] proposed three

modeling approaches: random network structures, hierarchical network structures, and scale-free network structures. These structures are represented as graphs. Each graph consists of a number of nodes which represent an individual of a network and are connected by edges which represent communication for exchanging information. Social science researchers usually have used random network structures for representing a social network in order to provide useful insights into dynamic and structural patterns. In this project, we intend to apply random network structures to model virtual social networks because they are more appropriate than hierarchical and network structures. Hierarchical network structures are well-suited for representing an organization, e.g., a company. Scale-free network structures have the property of having a few highly connected nodes and many with fewer connections and in a virtual social network such a pattern is rare.

Additionally, in this project we intend to use integration tools to determine group opinions in social networks. Consensus is an especially useful tool in this area. In the axiomatic approach to consensus theory [11], one of the postulates requires that the solution is the closest to all inputs. This is called optimality, as the task to solve is to determine the minimum of the sum of distances to all inputs. Depending on the distance measure used, different things may be determined. In this project this may be either the consensus opinion of the group, or the most influential member of the group. This presents an alternative, previously not researched, approach to social networks.

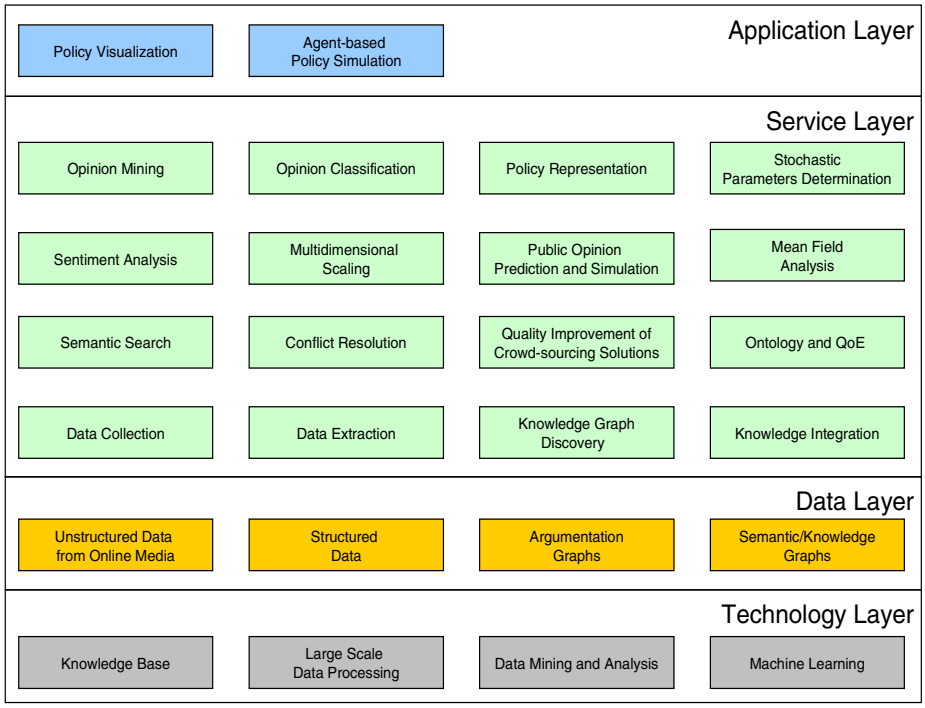

Fig. 3. The SOA-meta model and layers of the KNOWN software framework

3 Design of Software Framework

3.1 Software Architecture

Based on Service Oriented Software Architecture, we propose a SOA-meta model (see Fig. 3) for the KNOWN software framework that consists of the following layers:

- *Technology Layer*: The technologies which will be utilized by the system implementation.
- *Data Layer*: This layer contains some data structures and representations of information introduced and utilized by the project. These data will be stored in our cluster.
- *Service Layer*: The collection of algorithms and processes, which implements data building, manipulation and analysis tasks, and can be utilized as services by the applications developed in the project. Note that the results (methods, algorithms) of the KNOWN research will be implemented as services.
- *Application Layer*: This layer contains the applications developed by the project. These applications will use the services, and their operation can be defined using BPM [12] flowchart.

The relationship of the scientific and technical topics (to be presented in what follows) pursued in the KNOWN project is depicted in Fig.

The results (methods, algorithms) of the KNOWN research will be realized as services with well-defined interfaces. These services will be implemented as web services, using a proper software framework (for example: Java and the Spring Framework[1]).

For large scale data processing, MapReduce[2] developed by Google, with an open source implementation Apache Hadoop[3], or Spark[4] will be used in combination. Distributed databases, like Apache HBase[5], Apache Cassandra[6] (developed by Facebook), Voldemort[7] (developed by LinkedIn) or Dynamo[8] (developed by Amazon) will be considered for large scale data storage. For distributed processing and analysis of data stored in relational databases, there are several tools and frameworks like Pig[9] (developed by Yahoo!), Hive[10] (developed by Facebook) or Stratosphere[11], which is a generalization of the MapReduce framework optimized for processing big relational data. A distributed alternative to standalone statistical or machine learning frameworks is Apache Mahout[12], which can be used on Apache Hadoop distributed architecture.

---

[1] http://www.springsource.org/
[2] http://research.google.com/archive/mapreduce.html
[3] http://hadoop.apache.org/
[4] http://spark-project.org/
[5] http://hbase.apache.org/
[6] http://cassandra.apache.org/
[7] http://www.project-voldemort.com/voldemort/
[8] http://www.allthingsdistributed.com/2007/10/amazons_dynamo.html
[9] http://pig.apache.org/
[10] http://hive.apache.org/
[11] https://www.stratosphere.eu/
[12] http://mahout.apache.org/

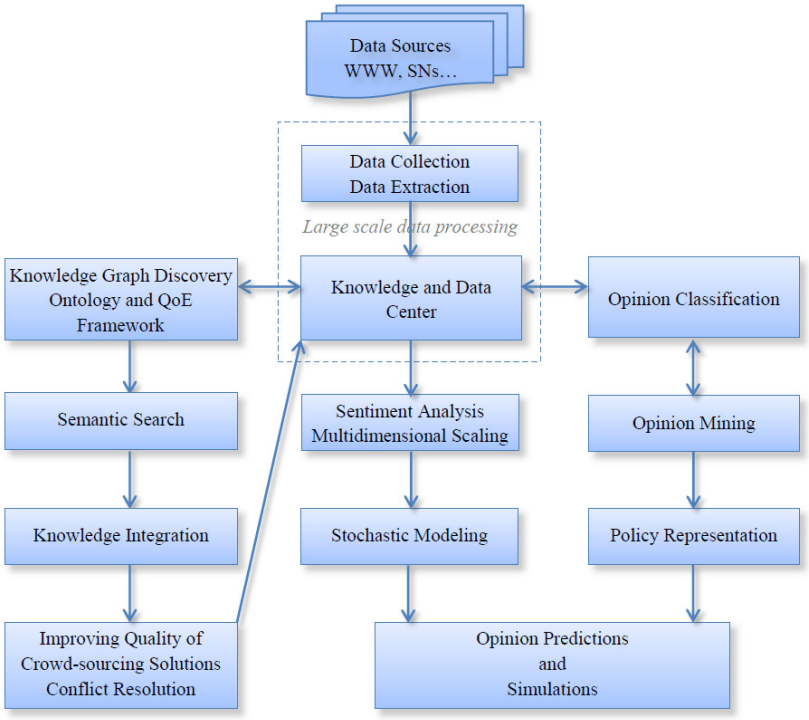

Fig. 4. The relationship of scientific and technical topics pursued in the KNOWN project

3.2 Applications

The applications of the KNOW Software Framework are interpreted as the orchestration of services that communicate through the defined interfaces and the jBPM integration mechanism[13]. The specific orchestration of services results in a process flows to satisfy a specific user need (e.g., to monitor the media appearances of policies issued by politicians/legislators and public opinions).

Information from focused data sources such as social networks, news portals and blogospheres will be exhaustively collected and extracted as raw data, which will be available in the Knowledge and Data Center. The center will be developed using advanced technology such as Hadoop and MapReduce to handle large information sources and keep up the system's performance. Further, raw data will be transformed by Knowledge Graph Discovery and Knowledge Integration (cf. Section 2.3) into structured data. In this process, the data will also be cleaned from inconsistencies by Conflict Resolution [11]. An exploitation view with interactions into data structures will be provided by Semantic Search. Ultimately, the structured and cleaned data will become available in the Knowledge and Data Center.

[13] http://docs.jboss.org/jbpm/

For the opinion predictions and simulations purpose, structured data will be processed further by the QoE and Crowd-Sourcing Framework, which defines a set of QoE metrics and quantitative measures of public opinions in general. The technique to explore distributional properties of citizens' opinions from the collected data and the method to map popular responses to policies, events and political processes will be done through Sentiment Analysis and Multidimensional Scaling techniques. Stochastic Modelling will simulate large-sized population projections, while traditional public opinion polling data will be collected on selected issues to validate the results obtained on a much more massive and cost-efficient way with our new tool. The output will be well visualized through the above mentioned opinion prediction and simulation.

Structured data can either be the sources for Opinion Mining and Classification, which is used as the input for Policy Simulation. The Policy Simulation uses graph-based policy representation to simulate reactions of citizens to policies based mined and analysed data. The mined and analysed opinions can also be used to predict and to simulate possible changes of citizens' opinions. These two simultaneous approaches of policy and opinion predictions provide complement results to each other because each technique can be used for different cases depending on the number of citizens.

In what follows, we provide an example for the use of the KNOWN framework to solve problems. Two scenarios are planned:

The purposes of scenario 1 are two-fold: 1) monitor new policies which will be issued by politicians and legislators and 2) trace public opinion and reaction to these new policy issues. For these purposes, we intend to develop a tool to visualize policies being issued and a tool for simulating public opinions (Fig. 5). The former tool can be developed through the tasks: 1) collection and extraction of unstructured data, 2) converting raw data into semantic and knowledge graphs, 3) transforming knowledge graphs into argumentation graphs which represent policies and as a result, policies can be visualized using a computer-supported argumentation tool. The development of the latter tool will require, in addition to the first two tasks, that semantic data in the form of knowledge/semantic graphs are analysed in order to mine and classify public opinion with respect to the policy being issued, and build agent-based simulation models. Deploying the developed simulation models we are able to simulate public responses to policies.

The second scenario aims at predicting public opinion with respect to a policy to be issued. As a new policy decision is introduced, it will be published in different media in relatively short time: news articles, blog posts, and social media commentary will appear online. We select and manually code in terms of objective characteristics (such as subject matter and the political actors involved) relevant content in a wide range of news media that cover the entire political spectrum of a country. Next, we trace the social media commentary (e.g. Facebook likes) that these news media products receive. Our new tools will be used to automatically collect and analyse the feedback of people to existing topics, providing input for policy simulation and opinion prediction (see Fig. 5). This is done automatically through scheduling tasks that run regularly and periodically. In particular, we will obtain measures of the intensity, valence,

over-time dynamics of public responses to a story, and can break these down according to the political sympathies and media preferences of social media users that can be determined from their publicly available postings in social media. This classification will be automated via our tool after a human-assisted learning period.

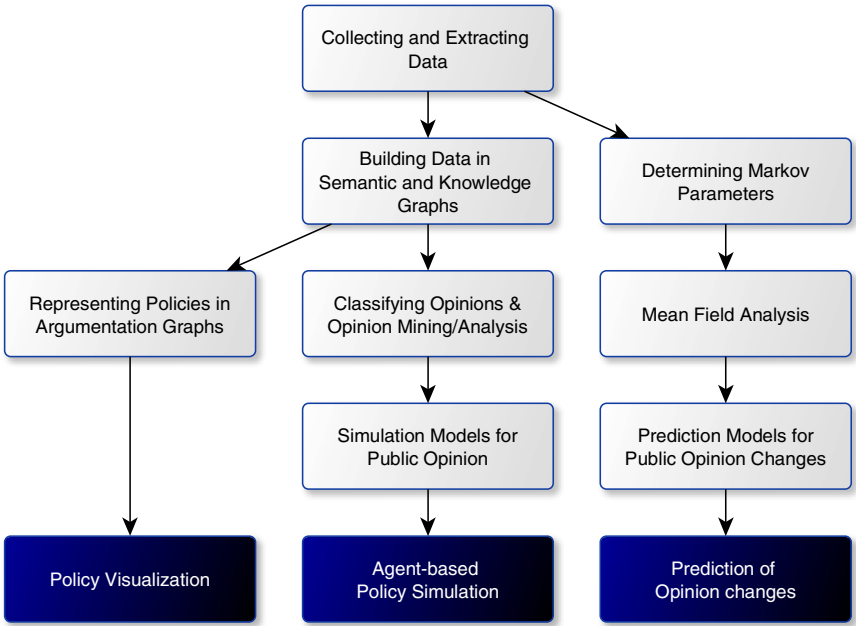

Fig. 5. Process-oriented description of the applications

With these data we also intend to develop a tool for predicting public opinion and possible changes (Fig. 5) because in a social network opinion leaders influence other participants in relatively predictable ways [4]. The development of this tool will be carried out through the following steps. First, we determine parameters for Markov chains and second, we use the mean field analysis and Multidimensional Scaling technique to map public opinion. In addition, we define a set of metrics which are referred to as Quality of Experience (QoE) measures that can be used to characterize the feelings and opinions of the public with respect to a policy issue.

4 Conclusions

This paper outlines the S&T challenges and methodologies used in the KNOWN initiative. At present, we are in the stage of the preparation of the project. We hope that our software tools can be implemented, which then can be used to monitor many measures related to public service provision and open government.

References

1. Bell, M.: Introduction to Service-Oriented Modelling. In: Service-Oriented Modelling: Service Analysis, Design, and Architecture. Wiley & Sons (2008)
2. Berners-Lee, T., et al.: The Semantic Web, pp. 29–37. Scientific American (2001)
3. Caracciolo, C., et al.: Results of the ontology alignment evaluation initiative 2008. In: Proc. of the 3rd International Workshop on Ontology Matching, held at the International Semantic Web Conference, pp. 73–119 (2008)
4. Huckfeldt, R., Johnson, P.E., Sprague, J.: Political Disagreement. The Survival of Diverse Opinions within Communication Networks. Cambridge University Press (2004)
5. Latane, B.: The psychology of social impact. Am. Psychol. 36, 343 (1981)
6. Lathrop, D., Ruma, L.: Open government: Collaboration, transparency, and participation in practice. O'Reilly Media (2010)
7. Lee, G., Kwak, Y.H.: An Open Government Maturity Model for social media-based public engagement. Government Information Quarterly 29, 492–503 (2012)
8. Manning, C.D., et al.: Introduction to Information Retrieval. Cambridge University Press (2008)
9. Mihalcea, R.F., Radev, D.R.: Graph-Based Natural Language Processing and Information Retrieval. Cambridge University Press, New York (2011)
10. Newman, M.: Networks: An Introduction. Oxford University Press, Inc., New York (2010)
11. Nguyen, N.T.: A Method for Ontology Conflict Resolution and Integration on Relation Level. Cybernetics and Systems 38(8), 781–797 (2007)
12. Scheuer, O., et al.: Computer-Supported Argumentation: A Review of the State-of-the-Art. Int. Journal of CSCL (2010)
13. Sommerville, I.: Software Engineering, 9/e edn. Addison-Wiley (2011)
14. Stocker, R., et al.: Network structures and agreement in social network simulations. Journal of Artificial societies and social simulation 5(4) (2002)
15. The White House. Memorandum for the heads of executive departments and agencies: Transparency and open government (2009)