Using Parliamentary Open Data to Improve Participation

Pierre Andrews
Instituto de Matemática e Estatística,
Universidade de São Paulo,
Brazil
pierre.andrews@gmail.com

Flávio Soares Corrêa da Silva
Instituto de Matemática e Estatística,
Universidade de São Paulo,
Brazil
fcs@ime.usp.br

ABSTRACT

For a lay-citizen, it is difficult to keep up to date with all the bills that are being discussed and will be voted at each level of the law making bodies of the government (e.g. a parliament, the congress, the senate). Currently, some citizens are very politically active and are taking part in the legislative process, starting petitions and protests to stop or support particular bills. However, the majority of the population is lost and does not care much for the details of the legislative process. They get politically active when journalists and news outlets raise an issue, but this is intrinsically biased towards the media's agenda and not always close to the citizen's personal interests. We are researching ways to automatically summarize and simplify the legislative process in a personalized manner for each citizen. In the way that Netflix or Amazon can learn a customer's preferences for movies, books and other products, we have been researching methods to build a system that can learn political preferences and topic of interest and can use these to automatically notify citizens when such topics are being discussed in parliament or when a bill is being voted that the citizen would strongly support or oppose. In this paper we present the state of the open data that is currently made available by legislative bodies to bootstrap such a system and we then discuss a particular use case, *Cidadão Automatico* that we are developing for monitoring the Brazilian legislative bodies.

Categories and Subject Descriptors

E.2 [**Data Storage Representations],** H.3.3 [**Information Search and Retrieval],** H.3.5 [**Online Information Services],** K.4.3 [**Organizational Impacts]**

General Terms

Management, Experimentation, Human Factors, Standardization

Keywords

Open Data; Open Government; Legislative Process; Parliament; Collaborative Filtering; Recommendation

1. INTRODUCTION

The transparency movement is trying to change the way governments work and interact with the public by pushing for the open access to all governmental data. While this idea might have seemed utopian a few years ago, the growth of the technologies surrounding the Internet has made the cheap distribution of data a fact. The public is now used to share content with their friends and to interact easily with the electronic world. Some governments are now seeing this as a way to improve democracy, fight corruption and extend the communication avenues with the citizens. This is actually pushed by a need from the public, which wants to have access to government services through the Internet [16], but also by a will of the government to decrease the cost of interaction with the public and create new value from data they have generated at public costs [13][5].

This movement has started by trying to reduce costs in existing, recurring, interactions between the citizens and their government: administrative requests. The governments have thus tried to reduce the costs of administrative procedures by moving them to automated online systems. These procedures have a very specific data flow, very close to the government as a platform concept introduced by O'Reilly [14]:

- Citizens fill in request forms – visa queries, document renewals or planning requests for instance – as "input".
- A government agency validates and accepts these requests.
- This agency then provides some sort of documentation as "output".

The logic of the transparency movement has used this early move towards online data sharing, to request data that are inherently public and that should be shared with citizens, journalists and other stakeholders in the society without going over red tape. Some also argue that openly distributing public data can create serendipitous opportunities and identify new value in the data or solve inefficiency issues in the government that wouldn't be detected internally [13][5]. A number of governments have thus voted laws to distribute such public data proactively, thus increasing transparency and, eventually, the democracy. However, such proactive distribution of data is different from the "input"/"output" systems that were fulfilling existing administrative needs. The data is now provided as "output" of the government without any specific request or explicit needs from the society [9]. This creates an issue, as the data is not always delivered in a way that is directly usable to the lay citizens. It is thus up to external entities, such as non-governmental organizations (NGO), to clean and structure the provided data and create value for the citizens [7].

Höchtl et al. [8] have argued that following the "Linked Open Data" (LOD) principles [10] is necessary to take advantage to the transparency data and guarantee the compatibility with external applications. Thus, while the long term goal of opening the governmental data is to let external entities create new value from the data, we argue that in the short term, governments should focus on standardization efforts for distributing their data following best practices of the web of linked entities (as exemplified in [3]). We show in this article that there is still a large gap between the way the data is provided and how the data might be used by external entities and then discuss how we could move past the data distribution issue towards systems to improve the participation of citizen in the democratic process. We first

ICEGOV2013, October 22-25, 2013, Seoul, Korea

study different sources of open legislative data, in Brazil, but also in the USA and in Europe and show that, while "open", they are missing many of the properties that would make them real Linked Open Data and make them of real use for external entities. We then introduce a citizen participation platform, based on a project that is being developed independently by the University of São Paulo, which would benefit from semantically linked open data describing the legislative process.

2. STATE OF PARLIAMENTARY DATA

Governments are creating a large amount of data and distributing diverse types of data to the public. In our research, we are interested in the open data effort surrounding the legislative process of different countries. The evolution of the legislations to novel issues and solutions is one of the core prerogatives of a government and therefore, one that is under scrutiny by citizens [16]. This is why governments are trying to improve the transparency of their legislative process [11], to improve the understanding and trust of the citizens in their own government's process. In our project, presented in the next section, we focus on improving the participation of Brazilian citizens in the Brazilian legislative process. For this use-case, we have thus gathered data from the Brazilian legislative bodies, and we have also looked at how the legislative transparency data was distributed by other countries. In this section we will discuss the existing open data effort of three governments and of linked NGO efforts: the United States of America (USA) government, the European parliament, the Brazilian governments and the Italian congress.

2.1 Data Scope of the Legislative Process

Before describing the data provided by these different countries, we need to define the scope of such data. Any government creates a very large amount of data and is divided in many bodies and branches with separate but linked prerogatives. In this article, we are focusing on the legislative branch of governments; that is, the branch that is responsible for writing and amending the laws enforced by a government. In federal states, such as Brazil or the USA, this legislative process is divided between the federal level and the state level, and often we can find a third, lower, level preparing laws at a municipal level. We are interested in countries where the law making process follows a legislative process, where a kind of deliberative assembly votes and amends laws; in the USA this takes the form of the bicameral US Congress while in some other countries it takes the form of a parliament.

The process of making a new law or amending a law can be abstracted to a generic process, even if details in each government might differ on particulars. The legislative process starts with a *law project* (sometimes called bill) that is proposed by some body. This body could be the executive part of the government, a member of the parliament or represent another source. For instance, the British and Swiss governments both have a process by which individual citizens can petition the public to push a new law project in the parliament.

The law project then goes through a preparation form, being studied by different *technical committees* that have particular expertise on the *topic* of the project. Law projects can be very broad and thus often go through multiple of these committees. Each committee is composed of a number of members that can influence the law project and sometimes vote at a committee level on the project.

Once the project has gone through the committees and is considered ready for discussing in the parliament, it is presented to all the members of the parliament that can discuss it, propose direct *amendments* and finally take a vote on writing the law project into a new final law. In bicameral legislatures, a law project might be voted by one of the two chambers of the parliament (in the USA, the "Senate" and the "House of Representatives") or pass through both chambers. An existing law can also be amended in a similar process.

This is a high level view of the legislative process that is already quite complex and where we can identify a number of entities and actions that are interlinked:

Law A law is an existing text of law, often made of *articles* that can be *amended*.

Law Project A law project is a text, often made of *articles* that describe a law to be voted.

Topic A law (project) is often attached to a number of topics.

Committee Each *law project* goes through a number of committees that will propose changes to the initial proposal and vote on the form of the project. A committee is composed of a number of members.

Representative A parliament has a number of voting members, distributed in two *chambers* for bicameral legislatures, which can take a number of actions on a *law project*.

Amendment Each law (project) can be amended to change part of its content.

Vote During the process of creation of a law or an amendment, representatives cast votes supporting or objecting the project.

We can see in Figure 1 that these entities are strongly connected, but in addition to these metadata level entities, which describe the creation process of a text of law, such a text also contains references to other entities, such as organizations, locations, or other existing laws that need to be linked in the full text.

From this introduction to the legislative process, we can see a number of entities that are directly related to the parliament, however, in our opinion the most interesting entity is the *representative* one, as it describes a person that has other interactions in the world than with the laws. Representatives are elected, either directly or indirectly, by the citizens, and to support a strong democracy, it is important to be able to follow their behavior: Citizens that have elected a representative might want to see if they are voting according to their own values. Such a feature is demonstrated in the United States of America (USA) by the Sunlight Foundation by its www.govtrack.us and www.opencongress.org sites where citizens can follow the lobbies and donations' influence on particular representatives and link this to their voting behavior.

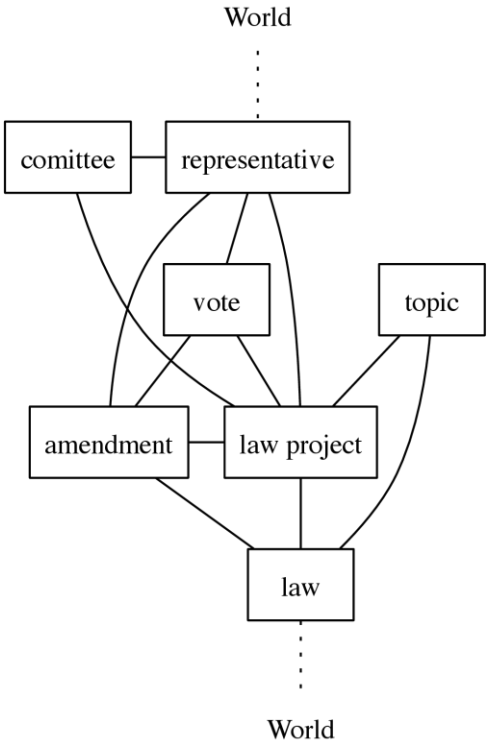

Figure 1: Interlinking between legislative entities

2.2 Legislative Open Data in the USA

The USA seem to be the most advanced of the governments we have studied in their distribution of open data, in particular with their data portal: data.gov. The advance of open data for the legislative process however, seems to be linked to NGOs that automatically process public data, from HTML sites, to extract structured data. The www.govtrack.us and www.opencongress.org sites have extracted data from the federal congress and the states' congresses so that citizens can follow the votes of their representatives. In a move towards open data, these sites are redistributing their data through a web application interface (API) that outputs JSON; the data is also available in a mix of XML with dedicated schemas and tab-separated files.

While the data is not available in RDF format, it follows most of the four principles of the Linked Open Data: it is available with dereferenced URIs through an HTTP interface, it follows a predefined schema and includes links to other related entities. However, the XML and JSON schemas used are not yet standards and are internal ad-hoc schemas. In particular, references to other entities (from law project to committees or representative) do not use the dereferenced URIs of these entities, but internal ids of the websites. rdfabout.com also provides an SPARQL/RDF endpoint for the *govtrack* data, but at the time of writing, the dereferenced URIs were not working properly. *govtrack* is explicitly cross-linking to entities representing the same representative on other websites, such as opensecret.org and vote-smart.org that allows for interlinking of the data and the creation of value. As an example, *influenceexplorer.com* can align the voting behavior of representatives with their "money trail", showing how lobbying and campaign finance is influencing particular bills.

The effort of these diverse NGOs is a great example of how interlinked data can be used to improve the visibility of the legislative process to the citizens.

2.3 Legislative Open Data in the European Parliament

In a similar manner to the USA government, the European (EU) parliament does not directly distribute open machine-readable data describing the actions of the parliament, but provides access to this information through the legislative observatory[1]. However, a private organization, Buhl & Rasmussen provides an API for automated access to this information thought the itsyourparliament.eu website. It is not clear how the data available from this website is obtained from the EU parliament, but it is made available through a web API returning JSON encoded data following an ad-hoc schema as illustrated in figure 2. We can see in the figure 2 that entities are not referenced by URIs, but with internal identifiers, in addition, the API is not dereferenceable as it requires an individual key for each person wanting to access it.

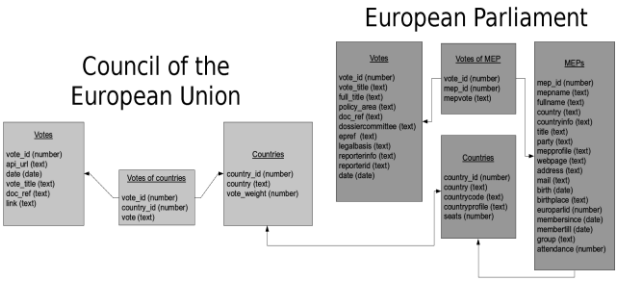

Figure 2: Extract of the *itsyourparliament.eu* simplified data schema, *source: Buhl & Rasmussen* api.epdb.eu

This schema contains external entities that could be dereferenced to standard datasets, in particular, the *country* entity is duplicated and refers to an internal entity not linked to the "external world". In the same way, the representatives (MEPs) are described by data embed within the ad-hoc schema and are not explicitly cross-linked to disambiguated entities in the LOD graph. This is of particular importance in the EU parliament as it is composed of representatives coming from diverse countries, where they have often been active in their local political life. Without being able to link them to external entities, it is difficult to follow how local politics influence their behavior in the EU parliament.

In addition, the EU parliament does not always vote laws directly, but votes directives that have to be then voted as particular law projects in each member countries. Therefore, having linked data for the EU parliament and each member country would allow to follow the evolution of EU directives in country specific legislative systems. We can see in this particular case that the interlinking and openness of legislative data is not just a technical issue, but would also, for example, benefit research in governance and policymaking.

[1] http://www.europarl.europa.eu/oeil/

2.4 Legislative Open Data in Italy

The Italian senate has adopted since February 2013 the Akoma Ntoso[2] XML format to distribute its legislative data. However, at the time of writing, it is still impossible to access this data in such a format.

The Akoma Ntoso format is an XML schema that has been developed by the African i-Parliament initiative, in partnership with the United Nations, and that is currently being standardized by the OASIS standard body into the LegalDocumentML format[3]. While it was first developed in Kenya, this format is now seeing a larger adoption, in particular with the European and Italian parliaments pledging to distribute their data with this format.

Currently, the Italian Legislative Data can be accessed with a custom RDF/OWL ontology[4] and through a SPARQL endpoint[5] in RDF format as well as CSV and JSON formats. However, at the time of writing, the URIs provided did not seem to be dereferenceable.

While, amongst the legislative bodies studied, the Italian senate seems to be more advanced in its choice of machine readable, structured formats to distribute the legislative data (including documents, votes, information about the representatives), we have not been able to evaluate the concrete use of these formats as the Italian's senate distribution platform does not yet seem to be fully accessible.

2.5 Legislative Open Data in Brazil

Brazil is a very large country with a federal structure. There are therefore multiple parliaments at different levels, and they are now pushed to deliver open data thanks to a new law enforcing proactive transparency. However, each government being independent, the process of opening data is not yet coordinated and we can find many different initiatives and distribution formats. In our research we are focusing on three particular parliaments that represent the three levels of government in Brazil: the Federal parliament, the parliament of the state of São Paulo (ALESP) and the parliament of the city of São Paulo. Contrary to the other case studies that we have discussed until now, each of these parliaments has already started, independently of each other, to distribute machine readable data, in addition to their frontend pages accessible to the citizens. However, as with the previous examples, there is a question of how standard and linked this data is.

2.5.1 Federal Parliament

The Federal parliament has a portal dedicated to open data and transparency at the federal level, the portal allows citizens to search for law projects, representatives and find information about the parliamentary process. However, the machine-readable data is still in a beta version and provides a subset of web-services[6], described through WSDL [4], returning XML results with ad-hoc schemas. These schemas do not seem to have documentation or to be described in any machine-readable format (XSD for instance).

```
<deputado>
  <ideCadastro>65498</ideCadastro>
  <matricula>356</matricula>
  <idParlamentar>8787554</idParlamentar>
  <nome>ANTONIO DA ...</nome>
  <nomeParlamentar>...
FERREIRA</nomeParlamentar>
  <sexo>masculino</sexo>
  <uf>MA</uf>
  <partido>PSC</partido>
  <gabinete>454</gabinete>
  <anexo>4</anexo>
  <fone>...</fone>
  <email>...</email>
  <comissoes>
    <titular>
      <comissao
        nome="Subcomiss\~ao Especial da Rio
+ 20"
        sigla="SUBRJ+20" />
    </titular>
    <suplente />
  </comissoes>
</deputado>
```

Figure 3: Sample XML describing a Representative in the Federal Parliament

If we look at the example provided in figure 3, we can see that it does not follow the LOD principles [10]: in addition to not following a standard format (or at least a documented format), the resources are identified by internal identifiers and sometimes (such as for the committees), the identifiers cannot be resolved to more information through the web-services. Another big limitation of the web-services is that there is no way to discover the existing entities, that is, there are web-services for searching based on a term, a particular identifier or a name, but no web-service is provided to list all the entities. Therefore, even if the government already provides a good set of web-services for obtaining machine readable data, we are still forced to use web page scrapping to extract information and to use heuristics to transform non semantic data and link entities.

2.5.2 State of São Paulo's Parliament

The parliament of the state of São Paulo (ALESP) also provides an advanced transparency portal where willing citizens can find information about their parliament and its activity. The portal also provides this data in machine-readable formats, in practice, RDF/Turtle [1] and CSV snapshots. Unlike the previous example, the ALESP portal tries to provide its data in a standard format, but falls short of following the LOD principles. As we can see in figure 4[7], the entities are derefenceable[8] but are not:

[2] http://www.akomantoso.org/

[3] https://www.oasis-open.org/committees/legaldocml/charter.php

[4] http://dati.senato.it/DatiSenato/browse/21

[5] http://dati.senato.it/DatiSenato/browse/23

[6] http://www2.camara.leg.br/transparencia/dados-abertos/dados-abertos-legislativo/dados-abertos-legislativo

[7] the lines have been shortened to fit in this article.

[8] even if they seem to have two identifiers.

Fully described The law project example we have provided shows that some information is missing (link to the committees responsible for the project, topics of the project, etc.) or is described in a non-semantic manner (for instance the year of proposal is hidden in the title attribute). In addition, we can see that even if there are two types of entities in the example, these are not explicitly typed and are not directly distinguishable without reading their descriptions and titles; which a machine cannot always do automatically.

Interlinked Entities such as the author of a proposition, or related institutions are not linked by a dereferenced URI but are described in full text, often with different identifiers (representatives sometimes have their title prefixed to their name, or groups are referred to as "John Robert et al.").

Law Project

```
<http://www.al.sp.gov.br/bases/projetos/projeto-de-lei.ttl#projeto182540>
      dbpprop:author "Archimedes ..." ;
      dbpprop:status "ARQUIVADO" ;
      dc:description "Denomina Prof Nercy Amelia Marteline Daher a EEPG (A) de Vila ..." ;
      dc:identifier "http://www.al.sp.gov.br/propositura?id=182540" ;
      dc:subject "ESCOLA, PIRAPORA DO BOM JESUS (MUNICIPIO), DENOMINAÃ‡AO, EEPG (A) DE VILA ..." ;
      dc:title "Projeto de lei 187/1980" .
```

Legislature Procedures

```
<http://www.al.sp.gov.br/bases/processos/processos.ttl#processo1011825>
      dbpprop:related "TRIBUNAL DE CONTAS DO ESTADO DE S.PAULO" ;
      dc:date "13/05/2011" ;
      dc:description "Of. GCRMC 604/2011 - TC 032156/026/07 - Julgou irregular o ..." ;
      dc:identifier "http://www.al.sp.gov.br/propositura?id=1011825" ;
      dc:title "Processo 2431/2011" ;
      dc:type "Processo de Contrato" .
```

Figure 4: Sample RDF descriptions from ALESP

While some of the missing data can be automatically scrapped from the frontend portal and linked with heuristics, some, such as the votes, are distributed in PDF documents, which are inherently difficult to process automatically.

2.5.3 Municipal Parliament of São Paulo

The municipal parliament of São Paulo also has an extensive portal where citizens can follow the legislative process, search for law projects, representatives, etc. While in other countries, municipalities have little influence on laws, the municipality of São Paulo represents over eleven million inhabitants and the laws made at the municipal level can influence a big proportion of the Brazilian population. The portal also provides some machine-readable data in an effort to provide open data. This data is provided following either an adhoc, undocumented, XML schema (for votes for instance) or a non-standard tab separated file format (which is documented). As for the federal parliament, while the data is distributed in an open manner, it is not yet interlinked or following the LOD principles, with issues similar to the ones described earlier.

This lack of semantic and of linked entities is an issue in a government structure such as the one in Brazil as it is difficult for a citizen to follow all the different levels of legislative process. An example issue created by this lack of links in the data would be the one of an interest group focusing on creation of movies; such a group would be interested in following and lobbying for new law projects to support the creation of movies. Its members would want to elect and lobby representatives that will support their cause, but they will also want to know about new law projects and amendments that might go against their needs and values. Because of the distributed organization of the government, such law projects might appear in municipalities (voting laws on who can film in the street for instance), in states (voting specific funding initiatives) or at the federal level (voting changes to the copyright law for instance). It is thus difficult for members across Brazil to follow all the laws voted in their diverse municipal parliaments, their state parliaments and the federal parliament. If the data of all these governments were distributed in a standard format and interlinked, members of such an interest group could create and follow a mashup of the data focused on their values.

Another example is the one of following the constancy of a representative in his/her support for particular values. Representatives will often start at a local political level, for instance, supporting issues valuable to their neighborhood in a municipal parliament, but will then move up the political levels, sometimes arriving at the federal level where they shape laws regarding the whole country. Supporters of such a representative would want to make sure that they voted for someone that actually supports their values, in action and not just in their speeches. They would also want to see if the representative they are choosing to vote for has a compatible history in his/her legislative actions, which is compatible with their values. Currently, because the entities representing a representative are often non-existent and just represented by a variation of their name in the dataset, it is already hard to follow them within one level of government. If they were represented with linkable, global entities, which are referred to in the same way in all datasets at all levels of parliament, it would be easier to build summaries of their voting history, infer their main values over time, etc.

2.6 The Cidadão Automatico Portal

The Pew "Government Online" report [16] has shown that, in the USA, 22% of adult internet users have tried to access information about the legislative process of their country. While many governments are opening their parliamentary data on web portals, giving access to law projects, bills, votes and representative contact details, independent websites such as www.govtrack.us, theyworkforyou.com, www.marsad.tn or itsyourparliament.eu are already giving access to aggregate data showing statistics on the behavior of parliamentary representatives in different countries. However, the data available is following the Internet trend, transforming a data drought into a data flood, with Internet users having access to too much data without a way to filter out and

easily search for the content they are interested in. The legislative process is often a complex process and citizens have issues in following what is going on in their parliament(s).

For instance, the municipal parliament of São Paulo provides a portal describing the legislative activities, with an access to all the law projects since 1948. However, there is no direct way to find out what projects are being currently discussed in the parliament or their current state. We believe that citizens are not interested in following all the laws projects, but only the ones that are relevant to them. For these projects, the citizens might want to express their opinion, often by contacting directly their representative, but also, in a world of social networks such as twitter or facebook, by raising the awareness of their friends and contacts on a particular issue being discussed in the parliament.

We can see this issue as the general issue of finding relevant content on the Internet. For instance, finding new movies that you would like to see or new books that you might want to read. Until recently, this process of curating new, interesting content to consume, was left to the specialized media, with journalists reviewing books, movies or products and consumers following these recommendations.

"Word of mouth" has also always been used, with friends recommending "products"[9] to consider. With the advent of the Web 2.0 and social networks, this word of mouth process is starting to replace the influence of journalists as the Internet has made it easier for people to aggregate the advices of their friends. While this works well for day-to-day topics, it has not yet completely spread to the participation in the legislative process.

We have performed an informal survey of citizens in São Paulo and have discovered that they currently discover "interesting" law projects being discussed in the government mostly through the lens of journalistic outlets (TV, printed media). Some also discover issues being discussed through discussions with more politically involved friends or through Facebook discussions. Avaaz'[10] community petitions are also a very powerful enabler in increasing the citizens' awareness of the issues being discussed in their governments.

If we look at commercial centered web platforms such as Amazon, Yelp or Netflix, we can see that they are a step further in the use of social network and of crowdsourcing to personalize customers' experience. They have pushed the idea of the electronic word of mouth even further, providing the aggregation of the public opinion at large, giving customers with summaries of the reviews of many like-minded people that they might not even know. This process is called "collaborative filtering" [15] and powers, for instance, the recommendations of Netflix, that can infer, based on the movies you liked and the movies other users liked, which movies you will want to watch in the future. Amazon uses the same principle to recommend new products and books based on the products that you have bought and rated previously.

Currently, citizens are limited in their access to the legislative process by the filter of journalists, who choose topics they think are important or that are sometimes pushed by lobbyists. However, these topics might not be the ones of interest for a particular citizen and can be politically biased. We believe that this is not particularly the fault of journalists (that cannot cover every possible topics) nor of the citizens, as they are faced with a number of barriers to understand and follow the legislative process.

We have illustrated in the previous section how governments are trying to improve the citizen's access to the legislative process, but even if the government adopts standard machine-readable formats, we would argue that there would still be many issues that would stop citizens to directly participate in the law making process.

The first issue that we have noted when studying the legislative data available in Brazil is an issue of data flooding and data distribution. Let us take again the example of a group of citizens interested in making movies, the federal government might pass new laws about copyright, taxes, wages of artists, etc. that will directly influence the movies industry. Locally, the government might also pass new law projects impacting their work (by forbidding filming in public parks for instance), but at a municipal level, the parliament will as often vote on projects of lesser interest for our example group (for instance, renaming streets, traffic planning, building planning, etc.). The first issue for our example group of moviemakers is thus to be able to find laws at every level of the government that interest them and might impact their trade and to filter out the laws that are not directly relevant.

We believe that the second main barrier is the one of understanding the content of laws and the general legislative process that will produce a law. Law projects, written in legal language (legalese) combined with the process of back and forth between technical committees, amendments of laws, etc. might seem Kafkaesque to lay citizens.

We are thus developing the *Cidadão Automatico*[11] portal to research new methods to improve the citizens' access to the legislative process and tackle these barriers. In particular, we are trying to develop methods to detect a citizen's values and classify new law projects in three classes: laws they would vote for, laws they would vote against, and ambivalent laws. We can see this as a recommendation problem [12], similar to the recommendation of new movies on Netflix or of products on Amazon. In *Cidadão Automatico*, users can vote for specific laws they are interested in (either against or for) or for particular topics that are of interest, the problem of recommending new laws becomes a problem of *collaborative filtering* [15], where we recommend new items based on the users' votes on previous items. We can base this recommendation on two methods:

- Using a similarity measure between topics or law projects, we can align new law projects to past projects on which the user has voted and infer the user's preference,
- Using the votes of other users of the system, we can find the similarity of a given user to the other users in the system and use aggregated votes from the most similar users.

While the issue of determining a user's interest in a law project can be described as a recommendation problem, the similarity measures discussed in the previous paragraph requires to know the features of law projects and of previous voters; this is where having a linked dataset of entities representing the parliament(s) is important. In particular, the users with which we could compare the citizen to determine his/her values would naturally be the members of parliament, as we are sure that they will vote on most law projects, we thus need to be able to link representatives with their votes on past law projects.

[9] In the general sense of the term.

[10] www.avaaz.org

[11] Automatic Citizen, in Portuguese.

In addition, while users of Amazon and Netflix are proactive in deciding to buy a new product (book, camera, etc) or watch a new movie, we do not expect lay citizens to be proactive in checking the upcoming law projects that *Cidadão Automatico* would recommend. Amazon and Netflix users will return on their own to see new recommendations but for *Cidadão Automatico*, we need to find a way to interest "passive" citizens. Hence, in addition to finding new techniques to filter and recommend law projects, we need to find a way to notify the citizen/users of the portal to take an active interest in a particular hot topic or law project.

Cidadão Automatico should be able to send an email to a particular user, and tell him that a law project requires a particular action, helping the citizen to contact a relevant representative or raise the interest of friends on social networks. We are thus also studying techniques to predict which one of the law projects of interest for citizens would also be worth an action on their side. We are currently considering that such law projects are the ones for which a citizen's representative (or group of representative) would not vote as the citizen would vote. That is, a representative would vote "Yes" to a project a citizen is against, and vice versa.

We therefore have to find a way to predict what representatives will vote on a new law project before they actually vote as well as predict citizens' votes on the same law project. If we are able to do this, we can find the discrepancy between representatives' expected votes and the citizens' expected votes and notify the latter that they should contact their representative and their friends to change the outcome of a vote.

We therefore need the portal to be as automatic and personalized as possible. To achieve this, we gather the open data, where it be in a machine readable format or not, from the portals of the different parliaments of Brazil[12] and aggregate them for display on the portal. From this data we can extract two types of bills:

- Bills that were already voted, for which we can extract individual votes of representatives.
- Bills that are still under discussion, either at the parliament level or in a sub-committee.

We have implemented two algorithms to automate the process described earlier. The first one uses the historical votes of representatives to predict what each of them will vote on the upcoming bills. While such a prediction might depend on many factors, we have found in preliminary experiments on historical data from the federal government of Brazil that we can predict votes of representatives with an accuracy between 70% and 85% when using standard machine learning approaches based on the content of the laws, the party of the representatives and their votes on past law projects.

Figure 5. Cidadão Automatico Bill Summary

[12] The study is currently limited to the Federal Parliament, the State of São Paulo Parliament and the Local Parliament of the city of São Paulo.

The second, personalization algorithm, consists in predicting the interests of the citizen user of the portal. To do this, we are currently gathering data from citizens. Each citizen, once logged in (with a Google or Facebook account) on the portal, can cast votes and see other's votes as shown in Figure 5. Cidadão Automatico Bill Summary:

1- The user can vote "against", "partly against", "neutral", "partly in favor" or "in favor" for an upcoming bill
2- For each bill, the user sees aggregate voting information for representatives/parties (as predicted by the algorithm discussed previously) and from his/her friends.
3- The user can use the sharing features to notify his/her social network about a particular issue.

As for product recommendation, if the citizen user provides enough data about his preferences by voting on laws that are of interest to him, we can then predict the votes that he will take in the future.

In addition, many law projects are very specific and written in legalese, they are thus hard to understand for the lay citizens with no legal background. We thus expect users of the portal to be more inclined to express their interest on higher-level topics. Our experiments on this issue are still at an early stage, but we believe that it will be difficult to accurately summarize and simplify legalese texts algorithmically. We are thus exploring the use of human computation and crowdsourcing to allow experts and motivated users of *Cidadão Automatico*, to provide labels and lay-summaries of law projects. We believe that we can use techniques inspired by wikis and question answering platforms to motivate users to generate new content that will help less expert citizens. We also believe that such data could be directly provided by the legislative bodies when distributing the data as it should be in their interest that law projects are well understood by citizens. However, we will need to find techniques to control for political bias in summaries of law projects.

While the goal of our research was initially the development of recommendation and filtering algorithms to help the citizens interact with the laws they might be interested in, we are currently faced with the issue of gathering up-to-date data from the different legislative bodies in Brazil and of automatically making sense and structuring this data. As we illustrated in the previous section, the current formats in which the governments (in Brazil, but also in other countries) distribute their open data does not yet allow to directly develop algorithms leveraging the graph of entities in the legislative data as they rarely follow the Linked Open Data principles; providing only raw data that has to be cleaned and disambiguated before being linked and leveraged for higher level applications. Therefore, we have currently only been able to develop the *Cidadão Automatico* portal at a small scale (the municipal parliament of São Paulo) as considerable manual work has to be performed to convert the data provided in a structure usable by our predictive algorithms.

3. CONCLUSION

While we have started our research on a problem of information retrieval and filtering, we have described in this article a lower level issue that has to be tackled before anyone can easily use the open government data to add value for the citizens and tackle issues of participation, policy making and governance.

In fact, in Brazil, the law for data transparency [6] explicitly states that:

- The data should be openly accessible to the public.

- The data should be accessible automatically to external systems, in open and structured machine readable formats.
- Such formats should be detailed and documented publicly.
- These requirements seem to be directly aligned with the definition of the web of linked entities [2] by following the Linked Open Data principles:
- Use universal identifiers (URIs), not internal identifiers from your database,
- Use dereferenceable links for the resources, accessible over HTTP,
- Describe your entities with useful information, in standard formats,
- Inter-link the resources

In the examples that we have provided in this article, we show that the governmental agencies are trying to distribute their data following these principles, but lack the know-how and resources to distribute data following the Linked Open Data principles. As this article discusses, this is not an issue present only for the Brazilian government but is also found in the EU parliament or the USA congress. We believe that an effort describing a standardized process for distributing open parliamentary data is required and that it is a global public issue, that diverse governments and NGOs seem to be trying to solve independently and internally when it could be solved at a global level as many aspect of the law making process is similar across countries. In particular, many of the entities represented in the data are global entities, such as countries, governance topic or persons that would benefit from being linked to the open web instead of being repeated in each dataset. Sites such as govtrack.us or influenceexplorer.com are already showing that a large graph of entities representing the governmental processes and their actors can be used to increase the transparency of the government. We believe that an international research effort, supported by experts of the semantic web, in collaboration with standardization bodies and governmental bodies, would greatly help the effort of transparency but also enable more participative platforms. In particular, such effort should not be internalized in each government or in private organizations, as this would defeat the serendipitous effect that is sought by opening the data at large. In our opinion, international conferences and public interest groups – such as the W3 Government Linked Data Working Group [17] – seem to be the best outlets to develop open formats to distribute open government data.

4. ACKNOWLEDGMENTS

This research is funded by Grant #2012/03795-1, São Paulo Research Foundation (FAPESP) and by the Open Source Software Research Centre at USP - NAPSoL.

5. REFERENCES

[1] D. Beckett and T. Berners-Lee. Turtle - terse RDF triple language, W3C note, 2008.

[2] T. Berners-Lee. Linked data-design issues. 2006.

[3] C. Böhm, M. Freitag, A. Heise, C. Lehmann, A. Mascher, F. Naumann, V. Ercegovac, M. Hernandez, P. Haase, and M. Schmidt. GovWILD: integrating open government data for transparency. In Proceedings of the 21st international conference companion on World Wide Web, pages 321–324. ACM, 2012.

[4] E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana. Web services description language (WSDL), W3C note, 2001.

[5] J. Cook. Economic issues in funding and supplying public sector information. 2009.

[6] P. da República. Lei no 12.527, de 18 de novembro de 2011., 2011. https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm.

[7] T. Davies. Open data, democracy and public sector reform. A look at open government data use from data.gov.uk., 2010.

[8] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, and J. Hendler. TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web, In Press, Accepted Manuscript, 2011.

[9] E. Dyson. Release 2.0: A design for living in the digital age. Broadway Books, 1997.

[10] T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. Synthesis Lectures on the Semantic Web Theory and Technology, 1(1):1–136, 2011.

[11] N. D. Institute, S. Foundation, and L. A. N. for Legislative Transparency. The declaration on parliamentary openness, 2012.

[12] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. Recommender systems: an introduction. Cambridge University Press, 2010.

[13] T. Jetzek, M. Avital, and N. Bjørn-Andersen. The value of open government data.

[14] T. O'Reilly. Government as a platform. Innovations: Technology, Governance, Globalization, 6(1):13–40, 2011.

[15] J. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. The adaptive web, pages 291–324, 2007.

[16] Smith. Government online. Pew Internet and American life project, 2010.

[17] W3C, W3 Government Linked Data Working Group, http://www.w3.org/2011/gld/