Making California Legislative Process Transparent

Foaad Khosmood, Alex Dekhtyar, Hisham Assal, Franz Kurfess, Joanna Snyder

California Polytechnic State University, San Luis Obispo

Executive Summary

This White Paper describes the research and design of the software system called Digital Democracy. The Digital Democracy system will provide its users: individuals with interest in public policy, journalists and media organizations, public interest and government watchdog groups, and lawmakers with direct access to videos and transcripts of State Legislative Committee hearings, as well as analytical tools enabling users of the system to conduct in-depth research of public policy issues and the attitudes of state legislators towards them.

The Digital Democracy system has been envisioned by Dr. Sam Blakeslee, former State Senator in the State of California, and the founder of the Institute for Advanced Technology and Public Policy (IATPP) at the California Polytechnic State University (Cal Poly). Through detailed conversations with Dr. Blakeslee and his staff, the authors of this paper: three full-time faculty members at the Computer Science department, and two staff members of Cal Poly's Collaborative Agent Design Research Center (CADRC), have developed the concept of the Digital Democracy system, conducted a preliminary feasibility study, and prepared high-level system design. This document provides a detailed description of the Digital Democracy system including its functionality and intended uses, an architecture design and an overview of individual components, a breakdown of the system design into phases, and the implementation plan.

We propose to design the Digital Democracy system at Cal Poly by engaging a group of graduate and undergraduate students in relevant research and software development activities under supervision of Cal Poly faculty. The system is broken into a number of stages, with each stage planned to take nine to twelve months (one academic year to one calendar year) of total time for a team of three to four faculty members and six to ten students split roughly in half between research and development. The initial system, developed over the course of the first three phases will primarily store, analyze and provide access to the legislative committee hearings for the California State Assembly and the California Senate. This system, referred to in the document as the Digital Democracy system, will serve as the model for building similar systems for other US states, federal legislature (the Congress), as well as for various local legislative bodies. This paper concentrates on designing and developing the Digital Democracy

system for California; transition to other legislative bodies is subject to future research and is not addressed directly in this paper. It should be noted that an earlier proof-of-concept version developed during Spring 2013 was initially called "OpenGov". Screen captures included in this White Paper retain the original name of the web-site. At the direction of the IATPP, the project and web-site are now called "Digital Democracy."

The first phase of the project involves building a robust on-line system that allows users to search full transcripts of legislative committee hearings, displays videos of selected hearings/hearing fragments in a content-rich environment. The system will provide "on-the-go" access to legislative transcripts and information about the speakers at the hearings, both the legislators and the witnesses/lobbyists/members of the public who testify at State Legislative proceedings. Additionally, the system will provide some basic analytical tools to allow the users to track the positions of individual lawmakers on specific issues, compare their stated positions and their actual votes, and cross-reference this information with the contributions to their political campaigns (obtained from an outside watchdog source like MapLight[1] and incorporated into the system).

Phases two and three of the project will extend the analytical functionality of the Digital Democracy system and extend the user experience beyond the web portal and into the realms of social networking and mobile computing.

At its core, the Digital Democracy system consists of three core components: (i) information import and integration component, that brings data from multiple sources, including existing public databases, as well as the committee hearings video transcription and annotation process developed in-house; (ii) knowledge discovery component: a variety of knowledge mining tools that work with the obtained data to discover important and interesting information, and feed the system's powerful search engine and (iii) the system's front end component which displays the information to the users, providing a convenient research environment for them.

Work on the Digital Democracy system involves intensive design and development, as well as significant amount of research in the area of knowledge mining, natural language processing, machine learning and pattern analysis. The implementation plan calls for creation, on each phase of the project, of balanced teams in which students working on the project are divided roughly evenly between the applied research activities: investigating which text and knowledge graph analysis methods lead to the most accurate results, and development activities: building the Digital Democracy system and concurrently incorporating the results of the research effort. We propose that students working on the project are hired part-time by IATPP, while also using the work on the project for their Masters theses and/or senior projects.

[1] http://maplight.org

The authors of the paper believe that the Digital Democracy project, as envisioned by Dr. Blakeslee and as outlined in this paper is an ambitious endeavor with very high upside, as it strives to fill an important gap in government transparency. At the same time, we are optimistic, that with proper resources, we will be able to organize effective research and development teams for each phase of the project, to produce high-quality, usable and efficient deliverables at the end of each phase.

1. Introduction

Over the past twenty years Internet has emerged as the key platform and the key contributor to shaping public policy, political debate and citizen participation in democratic processes. In United States, where Internet access is widespread[2], it provides millions of people access to shared knowledge and opinions that inform the worldview, decision-making and political activity of individuals.

Proponents of government transparency and citizen participation in public policy debates have long realized the importance of the Internet, and the World Wide Web specifically to their efforts. While much of the information about the operation of the U.S. government and state and local governments is available, it has been traditionally hard to obtain in forms that are easily digested and directly address the user's line of query.

This White Paper describes one such contribution to achieving transparency in governance and enhancing public awareness of and participation in the public policy debates: the Digital Democracy system designed to provide access to comprehensive public domain information about the lawmaking process at the level of a State legislature.

As is the case in other State Legislatures, legislative committee hearings form the core of the legislative process in the state of California. The committee hearings, during which State Legislators, invited witnesses, lobbyists and members of the public express their opinions, ask questions and share information, provide unique insight into the legislative process, the positions of individual members of the legislature, the participation of the interest groups in the public debate, and additional ancillary processes associated with bills becoming laws. Unfortunately, while most committee hearings are recorded and the recordings are nominally publically available, the way in which they are available is not conducive toward promoting transparency in legislature. Only in recent years have video/audio recordings of the committee hearings been available on-line. Prior to that, the only publicly available records of committee hearings were available as DVD recordings (and prior to that - as VHS tapes). The recordings themselves are multi-hour affairs with tens of speakers testifying, asking questions and making statements interspersed with ceremonial parts of the proceedings. None of the recordings are transcribed. The recordings contain a treasure trove of material about the legislative process, but in its current form this material is not accessible to the public in any meaningful way.

The Digital Democracy system is designed to bring direct access to the proceedings of the legislative process in the State of California by combining together the information about the

[2] According to the U.S. Census Bureau, in 2010, 80.23% of U.S. households reported Internet use (http://www.census.gov/compendia/statab/2012/tables/12s1156.pdf). In 2011, Internet use reached 95% among adults 18-29 years old, and 87% among adults 30-49 years old (http://www.census.gov/compendia/statab/2012/tables/12s1158.pdf)

legislative committee hearings, including videos and their fully searchable complete transcripts, information about the bills discussed in the hearings, and publicly available information about people: legislators, lobbyists, witnesses and members of the public, who participate in the hearings. The Digital Democracy system will link all this information together and will provide search and analytical functionality, allowing the users of the system to gain unprecedented insight in the lawmaking process in California.

This White Paper describes Digital Democracy as the system to be built and provides information about the research and development activities required to build it, together with initial projected estimates of the resources it will take.

The authors of this paper are faculty and staff at California Polytechnic State University, San Luis Obispo. From February through June of 2013, the authors, assisted by a team of graduate and undergraduate Cal Poly students worked with Dr. Sam Blakeslee, former California State Senator and Founding Director of the Cal Poly Institute for Advanced Technology and Public Policy (IATPP) on the conceptual model and the design of the Digital Democracy system. Through this preliminary study, the author team has arrived at the conclusion that building the Digital Democracy system as a research and development project at Cal Poly is feasible and achievable. We have determined that the Digital Democracy system concept gives rise to a number of interesting and important research projects, whose results will feed the development of the Digital Democracy software system. This paper, in parts that describe resources and personnel required and project phases and assumes that the development of the Digital Democracy system and all supplemental research activities occur on Cal Poly campus, and are conducted by Cal Poly students under the supervision of Cal Poly faculty.

2. Background and Related Work

The Presidential election campaign of 2008 demonstrated the power of the Internet and social media in increasing citizen participation and collecting public opinion. President Obama signed a memorandum on Transparency and Open Government, in which he defined three characteristics for government: transparent, participatory and collaborative. Transparency promotes accountability and provides information for citizens about what their government is doing. Participation by citizens enhances government effectiveness and improves the quality of its decisions. Collaboration actively engages Americans in the work of their government.

There are many types of efforts needed to make the vision of open government a reality. Every branch of government needs to provide the tools to make its work transparent, participatory and collaborative. For example, the executive branch has to demonstrate how the budget is developed and how money is spent. It can provide tools to solicit public opinion on supporting

certain types of research, or discuss priorities for the budget, e.g. education versus job development.

The legislative branch can offer transparency into the legislative process by providing easy access to discussions of proposed bills with clear view of the participants and their affiliations. It can offer participation by providing a medium for citizens to express their interest and views on any new proposed piece of legislation. Collaboration can be supported by giving citizens the tools to view, understand and participate in the legislative process. To date California governmental efforts to inform its citizens have fallen short. In April 2013 California received a "dismal grade" from the U.S. Public Interest Research Group for the level of transparency its government currently provides to the public.[3]

The concept of open government applies to all levels of government: federal, state and local. At the state level in California there are some efforts to support open government such the Cal Channel, which provides access to video recordings of state assembly and senate discussions of proposed bills. The *LegInfo* web site provides access to every piece of legislation in the state of California that was passed in the last 20 years. The California Common Sense (CACS) project aims at providing help in deciphering the large volume of data that is produced by the government and offer some analysis on what it means to the average citizen. CACS digs into government data, collects data on specific issues, performs analysis on the meaning of the collected data then presents the findings visually in graphs and charts for easy interpretation. CACS vision of transparency goes beyond access to data and into comprehension of the work of government, turning data into information. The *MapLight* service provides documentation of funding sources (e.g. campaign contributions) and voting records for legislators at the federal and state levels.

The current efforts, while providing access to vital information, do not inspire the average citizen to consume the information because the volume is overwhelming and the search tools are limited. The average citizen needs easy access to the information that interests them without the clutter of additional irrelevant discussions. The goal of the Digital Democracy project is to provide transparency into the California legislative process by digging into video recordings of legislative sessions and conducting analysis on their content.

3. Requirements and Technologies

[3] "California gets dismal grades from U.S. PIRG in Government transparency, "http://www.cafwd.org/reporting/entry/california-gets-dismal-grade-from-calpirg-in-government-transparency"

3.1 Features and Requirements

The intent of the Digital Democracy system is two-fold. First and foremost, Digital Democracy will provide direct, full, and unfettered access to the content of the State of California legislative hearings, complete with information on everyone who speaks at the hearings (legislators, lobbyists, witnesses, members of the public) and the subject matter of the hearings themselves (bills, policy agenda items). Digital Democracy will analyze the transcripts of the hearings, and will provide to its users the tools to track the stances and affinities of legislators, participation of the lobbying groups in legislative sessions, and the relationship between the discussions at the committee hearings and the changes to the underlying bills.

Additionally, while the Digital Democracy system described in this paper targets the legislative affairs in the State of California, its design and implementation will inform similar efforts directed at achieving transparency at other legislatures. The intent is to make the Digital Democracy system a model for how other similar systems are built.

We outline the core requirements that constitute the vision for the Digital Democracy system. The requirements are broken into several categories: users, data, functionality, security, and interaction. We outline each category below.

3.1.1 Users The Digital Democracy system will serve a number of different groups of users, providing basic services that are easy to use for all users, but also allowing users with more experience, and more time to conduct in-depth analysis. From more general to more specific, the categories of users are:

- **General public -** Primarily, the State of California residents and other individuals interested in the legislative process in the State of California.

- **Media -** Members of the media services (newspapers, news agencies, on-line news sites and communities). The Digital Democracy system will enable them to research individual committee hearings, the history of moving bills through the legislature and participation of interest groups in crafting the bills.

- **Interest groups/non-profit organizations -** Employees and affiliates of groups and organizations involved in public policy debate in the State of California will be able to use the Digital Democracy system to track the legislative process, assess the positions of individual legislators and political parties on a variety of issues, research the effects of witness testimony on legislation and lawmaker opinions and stances.

- **Public policy researchers and scholars -** Political scientists and sociologists will gain the ability to use the Digital Democracy system to analyze the state and the progression

of public policy debate on various issues within the state legislature, track the history of individual bills, and analyze the work of legislative committees and individual legislators.

- **Legislators and their staff -** Lawmakers and their staff will be able to utilize the Digital Democracy system for an easy to access track record of their activity in the committee hearings, as well as for analyzing the work of individual committees and legislative bodies on a variety of issues.

3.1.2 Data The Digital Democracy system will incorporate and integrate publically available data from a number of sources. We recognize two distinct categories of data sources in the Digital Democracy system: data sources unique to Digital Democracy, and already existing data sources.

Data Sources Unique to Digital Democracy - At this point we have identified two interconnected data sources that are unique to the Digital Democracy system:

- **Videos of California Legislature committee hearings -** While the videos themselves are publicly available, access to them is difficult. The raw footage contained in the videos is not annotated and little contextual information is provided. The videos, however, contain a treasure trove of information that can affect public policy debate in the State of California.

- **Transcripts of Committee hearings -** In addition to providing access to the committee hearings themselves, the Digital Democracy system will include full transcripts of the hearings, annotated with attribution of all direct speech to individual speakers. To our knowledge, this information, at present, is not available to the public. The Digital Democracy system will be the first source of this information.

Integration of Data from Publicly Available Data Sources - In addition to the video footage and the content of the video, Digital Democracy will provide information on the surrounding context: legislators and other individuals speaking at the hearings and legislative sessions, bills and their versions, as well as publicly available information about lobbyists and campaign contributions. This information will either come directly from the on-line resource providing the data, or will be collected and incorporated into the database that will support the Digital Democracy system. At this point, we have identified three such information sources.

- **LegInfo -** LegInfo[4] is the official web site for California's legislature. It contains information about the California State Senate and the California State Assembly, the Senators and members of the Assembly, Senate- and Assembly-originated bills, and their history, complete with legislative actions on the bills and committee and floor votes.

[4] http://www.leginfo.ca.gov/ and http://leginfo.legislature.ca.gov/.

LegInfo makes the CSV (Comma-separated values files) versions of its data tables publicly available. The Digital Democracy system will rely on LegInfo data, *imported directly into the Digital Democracy database* for the information about the following database entities:

- Information about individual members of the State Senate and the State Assembly, their terms in the legislature
- Information about Senate and Assembly committee membership for each legislative term
- Information about the bills considered in the Senate and the Assembly
- Schedule of committee hearings

- **California Secretary of State database of registered lobbyists -** The office of the Secretary of State for the State of California maintains an up-to-date database of registered lobbyists[5]. This information is very helpful in identifying speakers at the committee hearings who work for lobbying firms. As such, the Digital Democracy database will import from the Secretary of State database the information about individual registered lobbyists, the lobbying firms and information about the employments of registered lobbyists by the firms.

- **MapLight -** MapLight[6] is a website that aggregates information about political donations to Presidential, Congressional and State Legislative candidates. MapLight California[7] provides information about donations to electoral campaigns of the Senate and Assembly members, as well as donations to various other political campaigns within the state. This information is made available to the public on-line, and is also distributed in a form of a raw data set (in CSV, JSON or XML formats) and via a web service supporting a specially designed API. We do not plan to duplicate the MapLight database inside the Digital Democracy database. Rather, MapLight data on campaign contributions to legislators will be retrieved via the MapLight API and will be used to enhance the analytical facilities available in the Digital Democracy system.

3.1.3 Functionality The Digital Democracy system will provide its users access to video recordings and full transcripts of the California State legislature committee hearings. The system will put the committee hearings videos and transcripts in their proper context by:

- Identifying speakers and providing the available information regarding the speakers

- Providing background information on the bills the hearings to which the hearings are devoted

[5] http://cal-access.sos.ca.gov/lobbying/
[6] http://maplight.org/
[7] http://maplight.org/california

- Providing full-text search functionality allowing users to search for specific statements, keywords, topics and issues

- Providing analytical tools that help the users determine the stances and opinions of the hearings participants.

To reach this goal, the Digital Democracy system will support the following functionality, outlined here at a high level:

- **Search -** The core Digital Democracy use cases involve the user providing Digital Democracy with one or more search terms. The Digital Democracy system will search the available committee hearing transcripts and determine those occurrences that are relevant to the user search query. It will then display the information sorted in the order of perceived relevance and organized by information categories such as:
    - committee hearings
    - specific fragments of the hearings
    - specific hearing participants
    - bills
    - biographical pages

- **Browsing and viewing -** Digital Democracy will allow users access to the video and transcripts of all committee hearings stored in its database. The transcript will be tied to the video, allowing the users to find, either via the search functionality (see above), or by reading the transcript itself, fragments of hearings that are of interest to them, and be able to view the video of those specific fragments without the need to view the entire, often very long, video. The transcripts will also be annotated, providing background information on the bill being discussed, and the individual speakers participating in the hearing.

- **Analytics -** The analytical components of the Digital Democracy system provide another core added benefit. While a wide range of legislator score cards and voting guides exists for both US and state legislatures[8], the vast majority of them are based purely on the lawmakers' votes on the bills. The Digital Democracy system brings to light not just the votes lawmakers took on the bills, but also any statements they made during the committee hearings. The Digital Democracy system will be able to:

[8]For example, http://www.sacbee.com/votingrecord/, http://www.calaborfed.org/index.php/site/page/force_for_progress_legislative_scorecards, http://www.ecovote.org/scorecard, http://capitolresource.org/action-center/legislative-scorecard/, http://www.seniors.org/godocuserfiles/2413.29837_R2_Report.pdf, to name *just a few* for the State of California.

- ***Find all statements made by a specific lawmaker*** (as well as any other participant of the hearings) that is relevant to a specific topic.
- ***Analyze the sentiment*** of statements made by any participant of the hearings to determine if the statement supports the bill under discussion, opposes it, or is neutral to it.
- ***Analyze the history of statements and the voting record*** of a legislator (as well as any other hearings speaker) concerning a specific topic of public interest, and their stance of the issue.
- ***Analyze the committee hearings transcripts*** and determine any influence the discussion in the committee hearings had on the evolution of the bill under discussion.
- ***Analyze the stances of lawmakers with respect to the data about their donors;*** determine if there is any correlation[9] between the stances/votes of the lawmaker and the record of political contributions.

3.1.4 Security The system is expected to have reasonable security measures. While the data stored on the system is meant to be public, security risks exist with respect to denial of service attacks and intentional manipulation of data. This level of security is typically provided entirely by or in conjunction with the data hosting service provider. We have identified several scenarios of concern that we plan to address either with software solutions or a "best practices" guidelines for the eventual hosting service provider. Specifically:

- **Denial of Service (DOS) attacks -** These attacks are meant to overwhelm the Internet-based service in order to prevent ordinary citizens from accessing the content. DOS attacks can be greatly mitigated by providing system robustness, technological counter-measures, as well as timely involvement of law enforcement. In most cases, hosting service providers offer additional protection in form of hardware-based firewall and other security devices. These possibilities can be explored when signing contracts with service providers to host the Digital Democracy system.

- **Digital vandalism attempts -** Attempts at vandalism can occur for many reasons. Certain types of vandalism attacks have no relation to the content of the service (i.e. attacker does not care about what the site contains and simply wants to compromise the site for other reasons). Other cases can include malicious attempts at disinformation or defamation of certain individuals about whom information is available through Digital Democracy. The only difference is that attacks of the first kind typically strive to be

[9] We are careful in stating that the system can only observe correlation, not causation of the actions. Additionally, and correlation may be accompanied by causation in one of the two possible directions ((a) lawmakers get donations from a group because they support the group's causes, or (b) lawmakers support group's causes because they get donations from it). Such determination is outside of the scope of the Digital Democracy system.

noticed, while attacks of the second kind would prefer not to be noticed as to maximize their disinformation goals.

- In both types of cases, the same set of security prevention and data integrity techniques exist. These would be intrusion detection software deployed by a system administrator either hired by the project, or by the hosting service provider.
- In extreme cases, where risk of attack is substantial, data transport security mechanisms could be deployed in the system. The job of these mechanisms is to guarantee integrity of the data from its point of origin to the point of delivery. Typically encryption and parity checking is involved. However, in our assessment the risk does not warrant the extreme level of data insurance mechanisms.

3.1.5 Interaction Users will be able to interact with the Digital Democracy system through a number of channels. We briefly describe each interaction channel planned for the Digital Democracy system below.

- **Web site -** The entire functionality of the Digital Democracy system will be delivered to its users through the **web site portal,** which will be the first, and the primary interaction channel for the system. User interactions on the Digital Democracy web site will be optimized for three categories of use cases:
  - ***Individual lookups and quick searches*****:** i.e., short-form interactions of users with the web site of the "question asked - question answered" form.
  - ***Browsing and committee hearings viewing:*** The Digital Democracy website will provide convenient and easy-to-use functionality for browsing the content of the system. In particular, the Digital Democracy website will provide a rich and easy-to-use user experience for viewing the videos of committee hearings.
  - ***In-depth research:*** Prolonged studies of public policy issues and how they play out in the legislature supported by the in-depth analytical functionality of Digital Democracy.

- **Mobile applications -** A number of important user categories (media, public interest groups, and lawmakers) will have use cases for using Digital Democracy functionality in situations where their access to desktop or laptop computers is limited. A collection of Digital Democracy mobile applications targeting most popular mobile platforms (iOS, Android) and most popular mobile device formats (mobile smartphone, tablet) will be developed and deployed to allow interested users to gain access to Digital Democracy functionality on the go. The functionality of mobile applications will target the following categories of use cases:
  - ***Individual lookups and quick searches:*** helping users find the information they need quickly and present it in a convenient way on a (possibly) smaller screen.
  - ***Browsing and viewing short video fragments:*** *a convenient user experience for viewing committee hearings, geared towards viewing hearings in fragments.*

    - ***Short-form research:*** simple to express analytical requests.

- **Channels for sharing and social Networking** One of the core goals of Digital Democracy are increased legislative transparency. The Digital Democracy system will store much information, but this information will only have an impact on public discourse *when it is discovered and shared!* We approach the need to share discovered information in two ways:
    - **Sharing interface functionality -** Both the web site and the mobile applications will provide functionality for sharing discovered information (a snippet of a committee hearings video, a portion of a transcript, any result of the analytical tools - e.g., information about a lawmaker's stance on a specific issue) via popular social networks that provide sharing mechanisms. We expect the list of social networks and web sites for which the sharing functionality will be developed to evolve over time. However, at the outset, sharing via Facebook, Google+, Twitter and LinkedIn will be pursued.
    - **Plug-in functionality -** For social networks that allow third-party plug-in applications, Digital Democracy plug-ins reflecting the nature of the social networks and the expected user interaction experiences will be developed. For example, a Facebook plug-in for Digital Democracy allowing Facebook users to check what legislators from their districts say about specific issues will be developed as a pilot. Specific need for social networking plug-in applications and their functionality will be determined using the user requirements specification process outlined below.

3.2 Elicitation of User Requirements

This section outlines the requirements and proposed functionality for the Digital Democracy system at a very high level of granularity. The actual research and development for Digital Democracy will incorporate a clear and transparent process for determining specific user requirements, concentrating in particular on system functionality and usability. This process is briefly outlined below.

To some degree, features and requirements are derived from the needs and expectations of the intended user communities; we will refer to this set of requirements as User Interaction Requirements (UIRs). We are planning to use an agile development process that allows us to adjust UIRs based on feedback that reflects user expectations. The following methods will be used to elicit UIRs:

- **Expert suggestions and evaluations -** A small group of people with knowledge of the domain (the legislative process and related aspects), types of users (e.g. legislators, legislative staff, lobbyists, general public), and technical aspects of the proposed system will identify important use cases, extract core features, and suggest ways for different

users to interact with the system. This group of subject matter experts (SMEs) can be constituted as an advisory board that exists throughout the development cycle and possibly the lifetime of the system, or as an ad-hoc set of interactions with specific experts as needed.

- **Questionnaires -** These can be targeted at and distributed to potential user groups, requesting suggestions for features, or feedback on the system at various stages of the development process. With Web-based questionnaires, reasonably large groups of users can be reached with low overhead, and the analysis of their responses is fairly straightforward. However, the participants that actually respond are self-selected, and may not necessarily be representative of the actual user group. In addition, their responses may reflect desires, which may or may not be actual needs. Particularly in the early stages, superficial properties of the system may overshadow more critical features, and tilt feedback towards less important aspects.

- **Interviews and focus groups -** Individual users or small groups are involved in face-to-face or video conference conversations to discuss their desires, expectations, needs, and reactions to existing versions of the system. Conducting and evaluating these conversations is rather resource-intensive, but often yields insights not discovered earlier.

- **Usability evaluations at various development stages -** Experiments will be conducted with various versions of the project, ranging from sketches outlining very rough representations of the user interface and related interaction processes, over mock-ups displaying the appearance, and prototypes implementing partial functionality of the system. Participants will go through representative tasks involving typical scenarios for using the system. If possible, participants will be selected such that the major groups of users are represented (as captured in personas; see below). Again, these experiments can be time-consuming and expensive to conduct and evaluate, but deliver actual results based on users interacting with the system in a realistic manner, rather than their beliefs about features of the system.

The elicitation of UIRs will be used for the development documents below. It should be noted that the development documents will evolve in parallel with the above activities.

- **Use cases and usage scenarios** - Activities and tasks that intended users are expected to perform with the system.

- **Personas** - Hypothetical user archetypes that represent identified subgroups within user populations

- **Product features** - Related to user interaction

- **User Interaction Requirements** - Possibly with priorities

- **Evaluation criteria** - Related to user satisfaction and user experience; ideally objective and measurable, in practice often subjective and not directly measurable

In contrast to the other requirements, the UIRs have not been fully formulated at this stage since they will be derived from the information collected through the methods above.

4. Architectural Overview

The Digital Democacy system is designed as a knowledge management system dependent on several sources of information. Briefly in reference to the below Figure 1, the components on the left of the diagram are sources of information already available in various forms on the Internet. The far-right side of the diagram depicts methods of delivery of knowledge and interaction for end-users of this product.

Additionally, the system as envisioned uses a number of technologies, and depends on the successful use of a number of services:

- MAVIS transcription service
- External database sources: LegInfo, MapLight, Lobbyist Info from various sources
- Relational DBMS
- Web Server/web framework
- Social network plugins
- Mobile application enablers

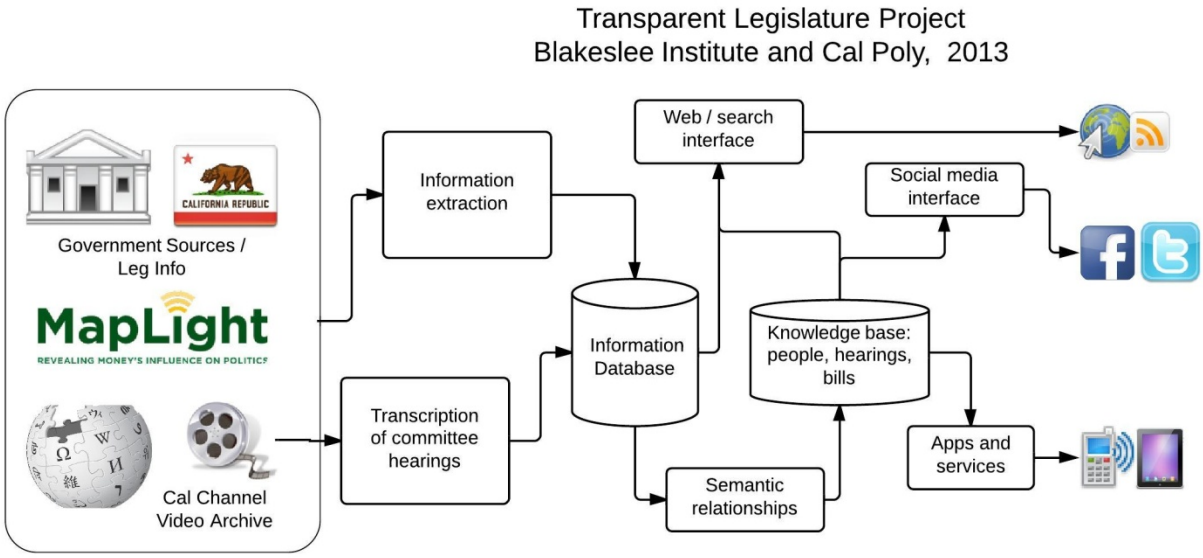

Figure 1

We collect information by directly interfacing with systems and websites that contain them. Of these, the Cal Channel video archive is most important since it contains video records of California legislature hearings that are not transcribed. Our system is designed to automatically extract content from the videos and contextualize this content with respect to persons

participating, bills/subjects under discussion and statements that were made. This knowledge is then accessible in various forms, notably a web interface with searching capabilities.

4.1 System Components: Overview

This section provides a brief overview of each component in Figure 1 and its relationship to the whole system. In the section that follows, more detail is provided regarding research findings, challenges and implementation.

- **Information Extraction** - This component consists of tools and scripts that could be invoked to transfer new information from a certain data source. Each of these scripts needs to be custom developed for each source of information. For example a particular script is needed to make a connection to the LegInfo server and download the latest LegInfo database, process it by extracting relevant data and copy them into corresponding table(s) of the Digital Democracy information database

- **Transcription -** This component performs exactly the same function as the Information Extraction component above. The only difference is this component specializes in video data. The component includes several steps: downloading the video/audio record of a certain hearing; passing that file to a video transcription tool like MAVIS; and finally annotating the resulting transcription by fixing known mistakes and recognizing known entities, and discourse signals. This latter part may require the assistance of an analyst. After some level of post-processing, the transcript information is stored in the Digital Democracy information database along with other content.

- **Information Database** - This component is used to store the information about the core entities represented in the system that has been obtained from various data sources.

- **Knowledgebase -** This component is also used for storage of data, but this data has been semantically analyzed and contextualized to some degree. In particular linkages and relationships have been found between the data from different sources. For example, a particular portion of video may have been transcribed as text (information), but further processed to tag specific persons, places, bills mentioned within it and perhaps linked to several other sources of information about those places or persons (knowledge).

- **Semantic Relationships** - This component is a catch-all label for many routines that can recognize semantic level relationships between various pieces of data. This component represents perhaps the most technically sophisticated part of the system, finding smart linkages and implications between raw data entries. The resulting knowledge is then stored in the Knowledgebase and feeds the analytical functionality of the system.

- **Web Interface -** This component is essentially a web server with search capability. It is able to accept user queries and return resulting knowledge from the Knowledgebase and the Information database and present them to the user on the web. The main design and development challenges of this component are user interface design and human computer interaction challenges.

- **Apps and Services -** This component provides an Application Programming Interface (API) or a programmable way to access the same content available in Digital Democracy databases. In addition access to the Digital Democracy Knowledgebase could be made available as a web based service. Using the API, developers can write utilities that allow interaction with the Digital Democracy system in different forms. For example, members of the public could download and place a "widget" on their own website which reports on the latest committee hearings that have been processed and their dates.

- **Social Media Interface -** This component allows Digital Democracy to first have a presence in various social media platforms such as Facebook, Google+ and Twitter. Secondly, interaction with the system, including setting up of automatic postings and alerts would be available for anyone to utilize. For example, a certain application could use the API to send alerts whenever new Digital Democracy information is published about a particular California lawmaker.

4.2 System Components: Details

4.2.1 Transcription - A major component of Digital Democracy is transcription of video/audio material. Ideally, an automatic process would create text versions of the media files and add them to the information database. Unfortunately, the present state of the science is inadequate for such a transcription system. The best automated transcription systems in the market could at best be described as "somewhat accurate". Appendix A outlines a two-term long student led project to obtain and evaluate the best performing transcription products. At the end of the process, Microsoft's MAVIS system was deemed to be the most appropriate for this project.

Although MAVIS performs transcriptions better than other evaluated systems as shown by our experiments, the resulting text is still far from accurate. Thus a post-processing verification phase involving a human editor is necessary for the foreseeable future. This phase serves two crucial purposes: First, the transcript itself becomes more accurate, and second, the kinds of errors that MAVIS makes become known and documented. The latter process helps our research team devise automatic corrections or suggestions that can reduce verification time for future transcriptions. We believe it is possible to reach a point where the level of automatic transcription accuracy is acceptable for this application.

4.2.2 Data Acquisition and the Digital Democracy Database - The Digital Democracy database will store information collected from a variety of sources (see Section 3.1.2), as well as directly from the hearings transcripts. The database will consist of two parts:

- **The data store -** The part of the database that stores the core data Digital Democracy operates with: hearings, transcripts, people, bills, committees.

- **The knowledgebase -** The part of the database that stores *the relationships between the spoken word in the hearings transcripts and other core database entities.* These relationships, for the most part will be established using the knowledge mining techniques described in the section below. Not all results of knowledge mining may make into the database itself, only those that constitute proper relationships between the entities stored in the database (e.g., a segment of a transcript mentioning a bill, or another portion of the transcript).

4.2.3 Nature of the database - At the evaluation stage of the project, we considered the issue of what database technology to use to host the Digital Democracy database. We have arrived to a preliminary decision to use *relational database technology* to build the Digital Democracy database. We have made this decision based on the following factors:

- The Digital Democracy database will import a large amount of data from outside sources. These sources (LegInfo database, Lobbyist database) store their data in relational databases, and where direct export of the data is allowed, supply the data broken down by relational tables.

- Most of the data in the Digital Democracy database is straightforward descriptions of entities and their relationships. Relational data model is a natural representation for such data. The key exception from this is the transcript text. However, there is a straightforward way (see below) to incorporate transcript text into the relational data model.

- There are multiple non-relational models that can potentially be used. Their advantages and disadvantages over the relational model and over each other need to be carefully studied. The preliminary study did not give us enough room to conduct such a detailed study.

- Relational DBMS such as MySQL can be used for prototyping purposes in an efficient manner.

At present, a prototype database design has been constructed and a prototype database built using MySQL as the DBMS. The Digital Democracy database stores information about the following entities.

The information outlined below will be sourced from the knowledge mining operations described in Section 5.

- **California Assembly and Senate Bills -** This includes information about bill versions, with the text of each version of the bill available, bill history and any votes taken on the bill. This information is sourced from the *LegInfo* database.

- **Assemblymen and Senators -** This includes the information about the terms of service for each lawmaker and their individual votes on bills. Sourced from *LegInfo.*

- **Assembly and Senate Committees -** Includes information about committee membership in each term. Sourced from *LegInfo.*

- **Committee Hearings -** Includes information about hearing schedules, bills discussed at each hearing, presence/absence of committee members at the hearings and the actual hearing transcripts with speaker attributions. Schedule and bill information is sourced from *LegInfo,* all other information is sourced from the committee hearing videos stored in and analyzed by the Digital Democracy system.

- **Hearing Speakers -** Information on all speakers at the hearings - legislators, lobbyists, witnesses, members of the public. The database will attempt to store the names and affiliations of all speakers[10], with additional information available for registered lobbyists. Lobbyist information is sourced from the *Secretary of State database of Lobbyists;* all other information is the result of transcript analysis.

- **Hearing Transcripts -** Each hearing transcript is divided into segments. Each segment includes a single utterance by one speaker at the hearing: a legislator, a witness, a lobbyist, or a member of the public. The database will also include information tying each fragment to the following other database entities:
    - mentions of bills
    - mentions of legislators, or other people participating in the hearings
    - mentions of specific overarching themes (e.g., education, budget, etc.)
    - mentions of any motions related to the bill
    - cross-references between different segments of the bill

As noted above, some important information featured in Digital Democracy, namely, the information about the donations received by the legislators will be sourced directly from *MapLight* without storing it in the Digital Democracy database. This preliminary decision is made

[10] Committee hearing rules specify that all speakers in front of the committee identify themselves at the beginning of their testimony; we will use analytical methods to automatically extract the information about speaker identity from the transcripts.

based on the fact that the *MapLight* data does not interact directly with the data on committee hearings that is unique to Digital Democracy. As such, the initial versions of the Digital Democracy system will interface with *MapLight* and bring its data into Digital Democracy dynamically for both the display and analytical purposes.

5. Knowledge Mining

The goal of knowledge mining is to identify and extract pieces of knowledge (sometimes referred to as “nuggets” or “knowlets”) from large data sets or collections of documents. In this context the assumption is that these repositories are available in a format that is accessible to computers and in volumes or formats that makes it impractical to be handled by humans. The terminology “automated” or “semi-automated” or “computer-assisted” knowledge mining is used to emphasize the share of work done by computers. One of the primary applications in our context is the processing of video recordings by converting audio information into text-based representations that reflect the content and if possible also the speaker of such recordings. Beyond that, knowledge mining techniques are used to enhance that information by adding metadata (such as time and place of the recording, duration, location of “snippets” within longer recordings, etc.) and context information (such as background information on speakers, related discussion items, etc.).

5.1 Overview

The assumption for the knowledge mining process is that there is a repository of machine-readable items of interest. This repository may consist of different types of sources (frequently databases, or text documents), may be distributed over multiple computers or locations, and may include access to repositories controlled by other organizations. Once an audio recording of a meeting has been transcribed and is available as a text document, our system will perform a natural language analysis, with Named Entity Recognition (NER) and Relation Extraction (RE) as core activities. NER identifies persons, objects, locations and similar entities that are mentioned in the transcript. Such entities can be referenced by their name (Jerry Brown, the Golden Gate Bridge, San Luis Obispo), a nickname or abbreviation, their function or title (the governor), or through relative references that have to be resolved within the context of sentences, paragraphs, or longer excerpts (such as “she”, “the speaker”, “the town”, etc.). Conceptually, entities referred to by their name are easiest to resolve; however, in practice there are numerous obstacles such as transcription errors, spelling variations, ambiguities (multiple persons or entities have the same name, or one single entity is referred to by multiple names, such as last name, first name, or initials). Frequently the extracted names can be validated against a list of known entities, such as the list of committee members or registered attendees for meetings. The other types of references such as functions, abbreviations or relative references can vary significantly in their difficulty to be resolved. The outcome of the NER phase is a list of recognized entities, if possible consolidated across names, functions, titles, abbreviations and relative references. In practice, some of the entries in this list will be tentative, expressed by a confidence factor in the correctness of the extraction.

The next phase, Relation Extraction (RE), examines the source document for relations between entities. This is often based on the familiar subject-verb-object pattern also used in sentences, as in “State Senator Mustang introduced Bill 123”. Beyond the extraction of information about

individual entities or relations between them, knowledge about the general topic addressed in a document, the opinions and sentiments of the authors (or speakers, in our case), and the extraction and rearrangement of specific statements (to summarize a document, for example) are of interest in our context, and will be examined. At this point it is unclear how successful their use in our context will be; this is part of the research to be done in the early stages of the project. There are successful usage scenarios for such approaches, but it remains to be seen how well they can be translated into text that is generated from spoken language and involves multiple speakers.

Natural language uses a large variety of words to describe entities and express relationships, and a simple extraction of the respective linguistic construct (noun or verb) may lead to an unmanageable number of terms. The reduction of the number of terms to a more manageable level is the goal of knowledge organization, often relying on classification and categorization of terms: Similar ones are grouped together, for example by replacing synonyms. A systematic approach to the organization of knowledge about a particular topic, or domain, utilizes ontologies. An ontology identifies the vocabulary used within a domain, and captures the relationships between various concepts, and between specific objects in the domain. Ontologies are often arranged around a hierarchical "backbone" of concepts; a good example is the biological taxonomy originally proposed by Linnaeus in his *Systema Naturae,* with kingdoms, classes, orders, families, genera and species. In addition, ontologies often perform the functions of dictionaries (definition of terms) and thesauri (synonyms and similar linguistic relations), concept maps or topic maps (more general relations between concepts and terms), and are sometimes visualized as knowledge graphs.

5.2 Natural Language Processing Technologies

There are several approaches to performing automated extraction of information using natural language text. These approaches often fall into one of two broad categories: rule-based approaches and statistical approaches. In this section, we will provide a short description of each approach and the benefits of each.

In a rule-based approach, an expert specifies exact rules, which a text phrase must match in order to capture an individual extraction. A domain expert knows what patterns are present in speech and how phrases are used in various contexts and the expert specifies these rules usually using some type of rule language. One of the primary advantages of a rule-based approach is that the extractions themselves can be typically very precise. Rule-based approaches tend to be difficult to maintain because of the fact that a human expert is needed to specify the rules and if additional rules are needed, new rules must be created. This makes transferring the information extraction process to and from different domain areas more work in order to create the new rules.

Statistical approaches on the other hand utilize the frequency of various phrasing and sentence patterns in order to build natural statistical models for information extraction. Given a known set

of parsed sentences or a set of labeled documents, various machine learning approaches can be applied to the data in order to build models of the parameters of the data. Given these models, predictions can be performed when a new sentence or document is provided. Statistical approaches can be more general and translate well from domain to domain, but the difficulty lies in providing labeled data sets to build the models.

5.3 Ontologies as a Means of Expressing Context

While computers can accomplish rudimentary processing of data (e.g., presentation, parsing, etc.) without an understanding of its meaning, software intended to intelligently analyze such content requires a representational model capable of expressing the concepts, entities, and relationships that together provide context. This can be accomplished utilizing a type of model known as an ontology. Ontologies provide an expressive, cohesive description of a domain that can be used by intelligent software components to answer in-depth questions as to the nature of relevant aspects of a domain. Typically rich with relationships, such domain knowledge is an integral component allowing intelligent software to effectively reason about an event or situation within the context in which it occurs.

Ontologies come in a variety of forms differing in their ability to precisely express intricate concepts and characteristics as well as their capacity to accommodate modifications in existing concepts or introduction of wholly new concepts while maintaining overall integrity and consistency. Well-crafted ontologies support these qualities through the exploitation of powerful analysis patterns [Fowler 1997] that engineer into such models qualities such as expression, flexibility, and extensibility.

The frameworks and tools used in this context will include the Web Ontology Language (OWL), Resource Description Framework (RDF), SPARQL Protocol and RDF Query Language (SPARQL). For all of these, various implementations are available, and the specific choices will be made at later stages.

6. Presentation

The main public interface of this project will be the searchable web interface. This interface is a website where a search bar features prominently. Entering keywords in this search bar and clicking search will return results in primarily three categories: people, bills and hearings. As noted earlier in this paper, the prototype application was initially called “OpenGov”. Following this initial phase of research and development and at the direction of the IATPP, the tool and project have been renamed to “Digital Democracy”. The screenshots were taken prior to the name changes. Future labeling of the application and project will reference the new name.

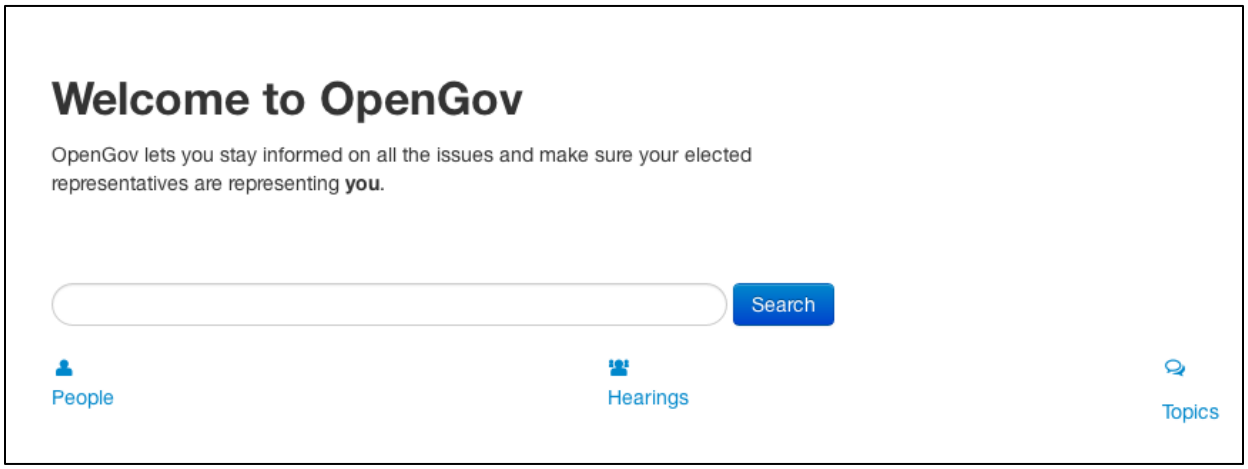

Figure 2

The initial use case for this interface is as follows:

1. The user enters one or more search term on the front page.

Figure 3

2. The user receives a page with several different kinds of results including hearing fragments, portions of person-pages and portions of bills.

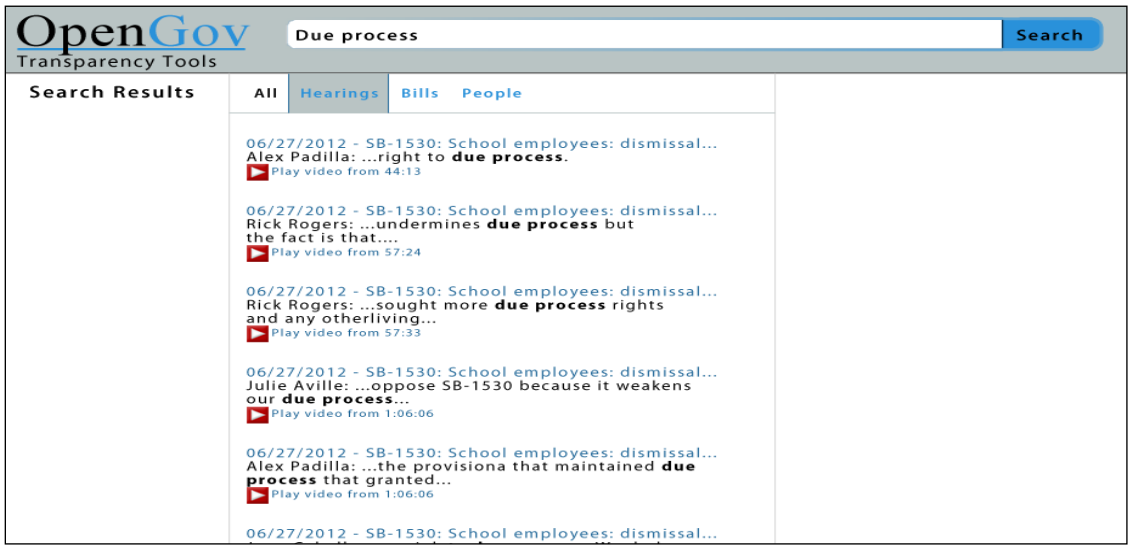

Figure 4

3. The user clicks on one of the search results.

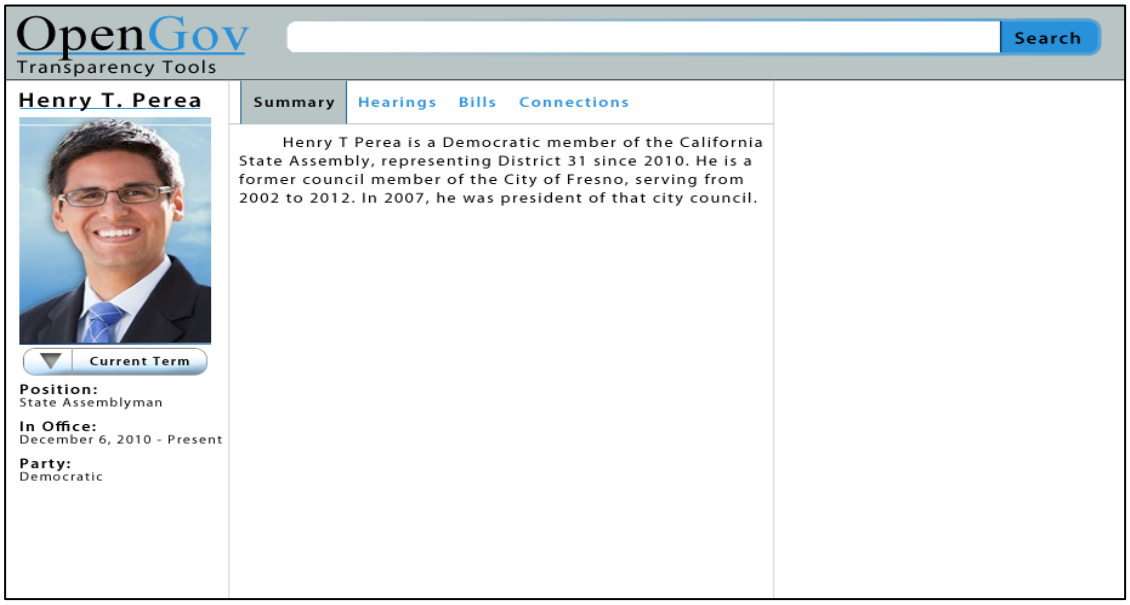

Figure 5

4. Hearing transcript results will also launch a video option

Figure 6

7. Social Media

The essential service of the system can be delivered through a variety of means besides just the web-based interface. Apps for mobile and hand-held devices can allow interaction with the system's knowledge-base similar to the web interface. Specialized apps can be created to allow delivery of particular types of information, for example, an automatic tweet or text message based on a search term appearing in a new transcript.

Similarly social media systems such as Facebook and Google+ can be incorporated into the system such that, for example, certain messages are automatically posted on them allowing for "likes" or comments within that space.

This portion of the project is subject to the greatest change and remains speculative. We estimate some amount of effort being required to write a typical social media app to interact with the system, but this estimate would need to be revised when we reach phases two and three of the project. We do know, however, that by making the knowledge base into an Internet service, the creation of many kinds of apps/social media plugins will be possible.

8. Cal Poly Capabilities

As discussed throughout this paper, this research was a collaborative effort among three Cal Poly entities, the Institute for Advanced Technology and Public Policy, the College of Engineering's (CENG) Computer Science Department and the Collaborative Agent Design Research Center (CADRC). Each brings its own distinguished background and experience.

8.1 Institutional Capabilities

Cal Poly is a nationally ranked, comprehensive, four-year public university with approximately 18,000 students. It is consistently recognized by "U.S. News and World Report" as one of the nation's top Universities primarily focused on undergraduate education, frequently the number one public institution in its category. It is designated as "Best in the West"[11] in the magazine's annual evaluation of public and private institutions providing undergraduate and master's level programs. Cal Poly focuses on the "learn by doing" philosophy. This philosophy, and the student skills that result from it, are frequently noted by recruiters as a pivotal reason why Cal Poly graduates are so highly sought by employers.

The Institute for Advanced Technology and Public Policy was founded by Dr. Blakeslee in 2012. Its mission is to advance student success by providing a unique academic environment that builds collaborative problem-solving skills while addressing real world public policy challenges. The Institute is self-supporting, non-partisan and committed to develop practical solutions to societal issues by informing and driving public policy through advanced technology. Dr. Blakeslee's vision is derived from a depth and breadth of experience as a scientist, business owner and legislator. His vision is central to the direction and pursuit of the Institute's goals.

Cal Poly's College of Engineering (CENG) has one of the ten largest undergraduate enrollments in the country. The College of Engineering was ranked eighth by U.S. News and World report among all public and private schools undergraduate engineering programs. Cal Poly's engineers and the skills they derive from the "learn by doing" philosophy are highly respected and recruited by industry leaders. Within CENG, the distinguished faculty of the Computer Science and Software Engineering Department has nurtured strong industry, corporate and

[11] http://colleges.usnews.rankingsandreviews.com/best-colleges/california-polytechnic-state-university-san-luis-obispo-1143

government partnerships which provide opportunities to students to participate in “real world” applied research projects.

Since its formation in 1986, the CADRC has focused on the design and development of intelligent decision-support systems, tools and services, with particular emphasis on contextual representation, multi-agent collaboration, modeling, simulation and visualization. The staff of the CADRC has over forty years of combined experience working in software systems research, development and project management. During that timeframe, CADRC has collaborated with faculty, students, administration, industry and government.

8.2 Curriculum Opportunities

The research described in this paper provides an example of student opportunity for research on “real world solutions”. To date, student research and development was either through a small stipend, course credit, Senior Project work or as part of their graduate research. This project provides flexible and beneficial opportunities for students to fulfill their intellectual passions while satisfying one or more of the following academic requirements.

8.2.1 Senior Projects

As a culmination of their undergraduate education, all Cal Poly undergraduates are required to submit for evaluation a “Senior Project”. The Senior Project integrates students’ theoretical and practical lessons into a “beyond the classroom” learning experience. Senior Projects consist of one or more of the following formats:

- a design or construction experience,
- an experiment,
- a self-guided study or research project,
- a presentation,
- a report based on internship, co-op, or service learning experience, and/or
- A public portfolio display or performance.

This experience and the documentation that follows are guided by Faculty and are normally related to the student's field of study, future employment, and/or scholastics goals. The goal of the Senior Project experience is to develop student’s ability to:

- Reduce a topic to specific points of analysis.
- Organize the points of analysis into a logical sequence.
- Apply acquired competencies to the successful completion of a project.
- Obtain, evaluate, synthesize, and apply project-related information.

- Develop and follow a project plan.
- Estimate hours of labor and/or cost of materials necessary to complete a project.
- Organize, illustrate, and write clear and concise project documentation.
- Accept supervision when needed.

8.2.2 Masters Theses

Department of Computer Science at Cal Poly offers an M.S. in Computer Science. While it is one of the smaller M.S. in Computer Science programs in the CSU system, at Cal Poly, this is one of the largest graduate programs. The M.S. program comes in two flavors. Cal Poly undergraduates with degree objectives in the College of Engineering can pursue a Blended B.S. +M.S. program which, upon completion of the requirements for both an undergraduate degree and the M.S. in Computer Science awards the students both degrees at the same time. The traditional M.S. program brings mostly students with undergraduate degrees outside of Cal Poly, or Cal Poly students with undergraduate degrees outside of College of Engineering. The M.S. program of study is very flexible: it consists of nine graduate and upper division technical electives and a year-long M.S. thesis study. Applicants with degrees outside Computer Science/Computer Engineering/Software Engineering can join the M.S. program, but are required to complete prerequisite coursework, consisting, depending on the background of the applicant, of one to twelve courses taken from the obligatory part of the B.S. in Computer Science curriculum.

Cal Poly M.S. students must complete a three quarter-long thesis. This approach exposes the majority of M.S. in Computer Science students to research, and allows the students to conduct a serious, year-long research project.

For the Digital Democracy project, we plan to primarily engage M.S. students to conduct applied research in the areas of knowledge mining, machine learning and natural language processing in order to solve the problems outlined in Section 4.3. We may also engage M.S. students early on to ensure that a single student stays with the project in multiple phases. Such students will be trained to assume some project leadership responsibilities during their second/third years on the team: develop and maintain the system's database model; develop and maintain the video-to-transcript conversion and subsequent annotation process; train incoming student members of the team in the processes employed on the project, and so on.

8.3 Faculty Capabilities

For the first few phases, the faculty members and affiliated personnel listed below have been directly involved in this project. Depending on the project needs and availability of faculty, others may join the team.

- Dr. Foaad Khosmood
- Dr. Alex Dekhtyar
- Dr. Hisham Assal
- Dr. Franz J. Kurfess
- Ms. Joanna Snyder

The background and capabilities of faculty and staff currently participating in the Digital Democracy project is outlined in a following section, “About the Authors”.

8.4 Facility Capabilities

The Digital Democracy project requires sophisticated computing and laboratory facilities to support research, development, deployment and subsequent growth and maintenance. The authors have secured capable resources for the foreseeable future.

8.4.1 CADRC Multi-blade, Large Capacity Server

Development and deployment of the Digital Democracy Prototype was completed on a high-end, seven blade Dell chassis. Future development, deployment and data-housing will continue on this server. The Dell chassis consists of 5 dual-Intel Xeon CPU (2.93GHz) with 96 GB of RAM, 1 dual-Intel Xeon CPU (2.67GHz) with 96 GB of RAM, and 1 quad-Intel Xeon CPU (2.00GHz) with 256 GB of RAM. These blades can be controlled remotely through the Dell Chassis Management Controller. It also includes a 7.3 TB EqualLogic PS6000 SAN and a Cisco 5500 ASA firewall router. The blades are configured to allow multiple environments for development, testing and deployment. This particular project is housed on a “LAMP” stack (Linux, Apache web server, MySQL DBMS and PHP). The space and speed of Dell system allows for high end computing and large data warehousing which will be particularly necessary as the project matures and integrates additional external data sources. The application is available from any location via a standard web browser interface.

8.4.2 Human-Computer Interaction (HCI) Lab

The HCI Lab consists of two rooms, separated by a one-way mirror. This setup allows the observation of participants engaged in interaction with computers or other devices, and is augmented by related technologies and tools. The primary purpose of the lab is to allow usability evaluations in a controlled environment, with facilities to record audio, video, keystrokes, mouse movements, and other interactions. In addition, it is suitable for interview and focus group sessions, where video and audio recording also can be very valuable.

9. Resourcing Digital Democracy Research

We propose to organize the work on the project in a number of phases. The first three phases are outlined below. During each phase, IATPP will engage a team of Cal Poly faculty to conduct

research activities and supervise project development. The faculty will determine the number of research and development student positions required for the project and will recruit the appropriate number of students to fill these positions. To ensure proper workflow and to provide IATPP and the Cal Poly faculty a level of control over the work of the students, we propose that the students are hired as part-time employees of IATPP, with assigned responsibilities determined by IATPP management and Cal Poly faculty based on the project needs. The students will be able to use their work on the project towards their required coursework: a senior project for undergraduate students (Computer Science, Software Engineering and Computer Engineering majors), or an M.S. thesis for the M.S. in Computer Science students. Such an arrangement will provide both a strong academic, as well as a financial motivation and will maximize the ability of Cal Poly faculty to engage the best and the brightest students in this project. Additionally, Cal Poly faculty working on this project may use portions of the research and development challenges that arise in the Digital Democracy project in the courses they teach.

Based on our analysis of our strengths and assets, the below section outlines required resourcing and technical goals for the current completed Prototype and subsequent versions resulting from financial and directional support. During the initial phases, a significant portion of efforts will be research related in order to discover and address the intriguing technical issues this project presents. As the initial phase demonstrated, the research and discovery process cannot be hastened proportionately to adding funding or people to the project.

Phase I (White Paper and Prototype) COMPLETED

Time: Spring-Fall 2013
Duration: 10 weeks
Staffing:

- 2 Faculty working 5 hours/week (partial gratis excluding payment for White Paper)
- 8 Students working approx. 5 hours/week (partial gratis or academic credit)
- 1 Project Coordinator working 5 hours/week (partial gratis)

Required Resources: 2 hours of MAVIS transcription service, $~33.00
Project Goals:

- Develop an application prototype able to be demonstrated as a use case (not deployable application) for prospective donors which incorporates a sub-set of applicable data sets.
- Identify technical challenges for further research and development.
- Identify potential tools required to be obtained or developed for future functionality gains
- White Paper outlining research methodology, findings and “way-ahead”.

Technical Goals:

- User accesses the website and enters one or more search terms on the front page
- User receives a page with several different types of results including hearing fragments, portions of person-pages and portions of bills results restricted to content from one hearing
- User clicks on one of the search results.

- For hearing transcript results there will also be a video launch option

Total Cost: $10,000

Phase II -- Alpha Version

Time: TBD (Contingent on funding)
Duration: 35 weeks (3 Academic Quarters)
Staffing:

- 1 Faculty (PI) working 10 hours/week
- 2 Faculty working 5 hours/week
- 9 Students working 10-20 hours/week
- 1 Project Coordinator working 15 hours per week

Required Resources:

- 35 hours of MAVIS transcription service, $875
- Miscellaneous supplies $125
- Web maintenance, back-up and security software $500
- Potential Data Subscription services $1000
- Travel $ 10,000

Project Goals:
A user-evaluated web-based application that allows access to an increased number (exact number to be determined) of committee hearings and the supporting knowledge graph

- Enhanced stability
- Meaningful progress and cost effective process to automate transcription, analysis and database storage.
- Incorporation of additional secondary data sets (ex., maplight.org and Leginfo)
- Define a repeatable process to communicate with external data sources
- Full developer level documentation of the system
- Support as requested for demonstration and publication documentation

Technical Goals:

- Semi-automated video-to-web process
    - web interface to be improved
    - development of new tool to facilitate human verification of transcripts
        - goals of this tool: toward a standard and accurate method of enhancing transcript data
        - should be web-accessible
- Enhance four sources of information into the system
    - videos
    - maplight
    - lobbyist database
    - leg info
- Automated import process for government data, other than transcripts
    - maplight
    - lobbyist db

- Clear manual process for transcript verification
- Establish System Administration process to maintain, update, back-up and secure web site
- Update and refine database schema / version control
- API specification for multiple entry points
- Research activities for phase II
    - User requirements elicitation
        - focus groups, interviews, questionnaires, subject matter expert advising
    - Usability evaluations
    - Natural Language Processing
        - Improve automated NL understanding
        - Knowledge graph generation
        - Begin work into semantic analysis with respect to stance detection, affinity for issues
- Phase III development plan based on phase II research (PI/faculty)
- UI development
- Documentation for new processes
    - adding and processing new video content
    - adding and processing new data sources
    - to be used to identify tools for automated transcription

Estimated Cost: $303,000

Phase III – Beta Version (Limited Release)

Time: TBD
Duration: 35 weeks (3 Academic Quarters)
Staffing:

- 1 Faculty (PI) working 10 hours/week
- 2 Faculty working 5 hours/week
- 9 Students working 10-20 hours/week
- 1 Project Coordinator working 15 hours per week

Required Resources:

- 50 hours of transcription service, $1,250
- Miscellaneous supplies $125
- Web maintenance, back-up and security software $500
- Potential Data Subscription services $1000
- Travel $ 10,000

Project Goals

- Semi-robust system which permits hardened testing throughout the system
- Beta application available for select end-users to evaluate and utilize
- Large scale server architecture that allows for exponential growth of the systems

- Full developer level documentation of the system
- Support as requested for demonstration and publication documentation

Technical Goals:

- Multiple entry points into system
    - Social media interaction (FB plug-ins, twitter)
    - Mobile - Phone and tablet applications for various systems
- API Implementation
- automatic transcription (improved accuracy)
- Evaluation of overall system design (database design)
- Evaluation of addressing increasing amounts of data
- Develop and stand-up system for full deployment, maintenance and versioning

Estimated Cost: $330,000

10. Conclusion

We have reached the conclusion that the Digital Democracy system can be built at Cal Poly by a team of graduate and undergraduate students working under the supervision of Cal Poly faculty and in tight cooperation with Dr. Blakeslee and the Institute for Advanced Technology and Public Policy. This decision is based on the following observations:

- The development of the Digital Democracy system ***requires a significant research effort directed at studying the following aspects of the project***:
    - methods for mining and analysis of hearing transcripts and extraction of useful information from them/tying the transcripts to other Digital Democracy data
    - user requirements analysis and usability studies

Cal Poly, with its stress on applied research, is well-positioned to meet these research challenges. Cal Poly faculty have sufficient expertise in working in the area of knowledge discovery from data and natural language processing, as well as expertise with supervising teams of students on long-term applied cross-disciplinary projects. Coursework taught at the Computer Science department exposes students to the technologies, algorithms and skills necessary for this project.

- The technical challenges associated with the software development for the Digital Democracy system can be solved using the knowledge of the faculty supervising the project and the skills of the students working on the project.

- The co-location of the Research and Development team and the project Customer (Dr. Blakeslee and IATPP) is conductive of active and seamless collaboration on the project.

- The research needs of the project dictate the timing of the project phases. While the development time can be sped up if Digital Democracy is developed on spec in the industry, because of the strict demands for research time, it will not be easy to exercise the time savings achieved this way.

- When completed via the mix of senior projects, M.S. theses and part-time on-campus employment at IATPP and supervised by Cal Poly faculty, this project will cost significantly less than if performed by full-time professionals in the industry.

About the Authors

Dr. Foaad Khosmood (Principal Investigator) is Forbes Professor of Computer Engineering and Assistant Professor of Computer Science at California Polytechnic State University. Dr. Khosmood has industry experience as a senior software engineer for Intel and graduate internships at IBM and Symantec. He is the lead author for an intelligent security device patent titled "offline packet analysis". Dr. Khosmood's research involves computational stylistics, natural language processing and artificial intelligence. He received his Ph.D. in computer science from the University of California Santa Cruz in 2011.

Dr. Alex Dekhtyar is a Professor at the Department of Computer Science, California Polytechnic State University at San Luis Obispo. Dr. Dekhtyar received a Ph.D. in Computer Science from the University of Maryland at College Park. At Cal Poly, he teaches primarily database courses, as well as some of the courses in the areas of data mining and bioinformatics that he has developed. His research interests include uncertain reasoning and management of uncertain data, semi-structured data management, software traceability, data mining and machine learning and bioinformatics. Throughout his 13-year old career in academia, which spanned two institutions (Cal Poly and University of Kentucky), Dr. Dekhtyar has worked on numerous cross-disciplinary projects that involved anthropologists, political scientists and architects, Old English scholars, irrigation engineers, animal scientists, and microbiologists and geneticists. Dr. Dekhtyar was a co-Principal Investigator on three NSF grants, and on numerous research grants from NASA. He co-authored 18 journal articles and 60 peer-reviewed conference and workshop proceedings papers, and contributed three book chapters to two different books. In 2013, his paper on automating software traceability published in the Proceedings of the International Conference on Requirements Engineering in 2003 won the "Most Influential Paper" award.

Dr. Hisham Assal is a researcher at the Collaborative Agent Design Research Center (CADRC) at Cal Poly. He received his Ph.D. in Computational Design from the University of California, Los Angeles (UCLA) in 1996 and joined the CADRC the same year. Dr. Assal designed, developed and supervised development of distributed, collaborative decision-support systems for military and industry applications. His interests include knowledge representation, knowledge management, intelligent agents, cyber security and complex systems. Dr. Assal published many papers in the areas of software development, knowledge representation and knowledge management. His most recent software project was a platform for knowledge extraction from natural language text, utilizing ontology, intelligent agents and NLP techniques. The results of this project were published in ACM and IEEE conferences and appeared as a book chapter in 2013. Dr. Assal's experience covers a wide range of topics dealing with complex integrated systems. He worked on Ontology-based systems, intelligent agents, Business Process Management (BPM), knowledge management, and supply chain management. The areas of application for these systems included logistics and planning, transportation management, and air traffic control.

Dr. Franz Kurfess is a professor in the Computer Science Department, California Polytechnic State University, San Luis Obispo, where he teaches mostly courses in Artificial Intelligence, Knowledge-Based Systems, Human-Computer Interaction, and User-Centered Design. Before joining Cal Poly, he was with Concordia University in Montreal, Canada, the New Jersey Institute of Technology, the University of Ulm, Germany, the International Computer Science Institute in Berkeley, CA, and the Technical University in Munich, where he obtained his M.S. and Ph.D. in Computer Science. His main areas of research are Artificial Intelligence and User-Centered Design, with particular interest in the usability and interaction aspects of knowledge-intensive systems.

Joanna Snyder is a retired Staff Emeritus from the California Polytechnic State University in San Luis Obispo. She continues to consult for Cal Poly's Office of Research reporting directly to the Vice President for Research and Economic Development. Prior to her retirement, Ms. Snyder was the Director of Operations of the Collaborative Agent Design Research Center (CADRC). Throughout her career, she coordinated multiple software development, testing and maintenance teams, including a Department of Defense (DoD) software system fielded to U.S. military sites world-wide and supported "24/7". This system was cited in the 2012 Congressional Budget Report as a positive example of a project that effectively fulfilled DoD Core Business Mission Results. Ms. Snyder earned her Bachelors of Science in Political Economy of Natural Resources from the University of California, Berkeley in 1984. To enhance her work experience, she has taken several units of Masters of Science coursework in Computer Science from the California Polytechnic State University. Additionally, she earned the Department of Defense Information Assurance Security Officer (IASO) certification.

Additional Reading

Gupta, M., & Jana, D. (2003). E-government evaluation: A framework and case study. *Government Information Quarterly*, *20*(4), 365-387.

Meijer, A. J. (2003). Transparent government: Parliamentary and legal accountability in an information age. *Information Polity*, *8*(1), 67-78.

Radics, A. (2001). Cristal: A Tool for Transparent Government in Argentina. *World Bank: http://www1. worldbank. org/publicsector/egov/cristal_cs. htm*.

Grimmelikhuijsen, S. (2009). Do transparent government agencies strengthen trust?. *Information Polity*, *14*(3), 173-186.

Kim, P. S., Halligan, J., Cho, N., Oh, C. H., & Eikenberry, A. M. (2005). Toward participatory and transparent governance: report on the Sixth Global Forum on Reinventing Government. *Public Administration Review*, *65*(6), 646-654.

Ndou, V. (2004). E-government for developing countries: opportunities and challenges. *The Electronic Journal of Information Systems in Developing Countries*, *18*.

O'Hara, K. (2011). Transparent government, not transparent citizens: a report on privacy and transparency for the Cabinet Office.

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, *27*(3), 264-271.

LIU, Z., & LIN, L. (2009). Transparent Government: A History and Logic View on the Transformation of Government Pattern [J]. *Journal of Sichuan University (Humanities & Social Sciences)*, *1*, 002.

Wong, W., & Welch, E. (2004). Does E□Government Promote Accountability? A Comparative Analysis of Website Openness and Government Accountability. *Governance*, *17*(2), 275-297.

Mulgan, R. (2000). 'Accountability': An Ever Expanding Concept?. *Public administration*, *78*(3), 555-573.Persson, T., Roland, G., & Tabellini, G. (1997). Separation of powers and political accountability. *The Quarterly Journal of Economics*, *112*(4), 1163-1202.

Lowry, R. C., Alt, J. E., & Ferree, K. E. (1998). Fiscal policy outcomes and electoral accountability in American states. *American Political Science Review*, 759-774.

Pasch, G. (2003). Convergencia de Contenidos: el Control de Objetos" rich media.

Stewart, Q., & Todd, D. (2009). Using university collections in digital library education. *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM.

Kirkland, J., Smith, A., & Roy, L. (2010). Capturing our stories and GLIFOS: rich-media video. *Electronic Library, The*, *28*(5), 650-656.

Hutchens, C. (2011). From Picas to Pixels: Innovative Oral Histories on the Web: An Interview with Quinn Stewart, Coordinator of Information, Technology and Lecturer at the University of Texas at Austin School of Information. *Serials Review*, *37*(3), 207-211.

Marroquín, G. XML saves the day: Porting a Rich-Media Collection to a Mobile Platform in Three Weeks Flat.

Krötzsch, M., Vrandečić, D., & Völkel, M. (2006). Semantic mediawiki. *The Semantic Web-ISWC 2006*, 935-942.

Buffa, M., Gandon, F., Ereteo, G., Sander, P., & Faron, C. (2008). SweetWiki: A semantic wiki. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(1), 84-97.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web*, 722-735.

Krötzsch, M., Vrandečić, D., Völkel, M., Haller, H., & Studer, R. (2007). Semantic wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(4), 251-261.

Appendix A: Transcription tool selection

Automatic transcription of audio and video recordings is both an essential part of this project, and a technically unsolved problem. The best automatic transcription technologies are still not producing accurate text.

Given the constraints of this project, we set out to evaluate the best transcription tool using a sample audio from a California Senate hearing.

Specifically, two experiments were conducted. First, we compared two freely available technologies: YouTube and CMUSphinx. In the second experiment we compared several different free and for-pay systems using the same standards. We concluded that MAVIS by Microsoft was the most accurate of all systems. The details of the experiments are summarized below.

Experiment 1: YouTube versus CMU Sphinx

This experiment evaluated two tools: YouTube's automatic closed captioning service and the CMUSphinx. The metric used to evaluate performance was word error rate (WER), word accuracy and the time it takes to retrieve the transcript. WER uses the Levenshtein Distance to compute what changes need to be made to a string of words in order to convert it to another string of words. The modifications are insertions (I), deletions (D) and substitutions (S).

Let N be the number of words in the correct string of words:

$$WER = (S + D + I) / N$$

Accuracy is the measure of how often the remaining words (not deleted or substituted) match the correct target word after transcription.

$$Accuracy = (N - D - S) / N$$

A low WER and high accuracy are desired. We used an 8-minute audio file taken from the CalChannel website to conduct these tests. We manually transcribed this audio clip to have a 100-percent accurate reference from which we could evaluate how YouTube and the CMUSphinx performed.

YouTube
YouTube offers a service for uploaded videos to be automatically transcribed. This tool is extremely easy to use, however, we lose control over the transcription process and all tuning

capability. The process is simple: upload a video to YouTube and when playing it back, select the closed-captioning button to view captions that were generated automatically. These captions can also be downloaded.

Table 1. YouTube advantages and drawbacks

Advantages	Drawbacks
● Free ● All processing is done by YouTube ● No configuration necessary (easy to use) ● API available to automatically upload new videos ● Time-stamped	● Undefined amount of time before transcript / closed- captioning becomes available (could be weeks) ● No control over transcription process

We uploaded the sample audio file to YouTube and about 24 hours later, the closed captioning service became available. However, another sample file of the same length was uploaded at a different time, and closed captioning did not become available for about a week. Thus, the amount of time needed to transcribe a file with this service is undefined. The WER was 0.29 and the accuracy was 0.76. From a human perspective, the transcript was readable, however the context of what took place in the actual audio clip was difficult to recover.

CMUSphinx

The CMUSphinx is an open-source speech recognition toolkit. As the name implies, this toolkit was developed at Carnegie Mellon University, and was first released to the public in 2000. The entire code-base is freely available under a BSD-style license. There are several packages included in this toolkit, each with different goals and objectives (speed, accuracy, mobile vs. server environments).

The package we evaluated was Sphinx-4, which is the most adjustable and configurable speech recognizer they offer. This package is written entirely in Java and is ideal for server / cloud environments due to its required resources. We chose this package because of its configurability, with hopes of tuning it specifically for government-related meetings to achieve higher accuracy.

Table 2. Sphinx advantages and drawbacks

Advantages	Drawbacks
● Free ● Configurable: Sphinx-4 is configured	● Configuration is complex and it is difficult to achieve high accuracy

with an XML file given at runtime. This configuration file defines many parameters that are used in determining how an audio file is transcribed. The three major parameters are the language model, the acoustic model and the dictionary. The language model defines several probabilities and statistics on words and their probabilities of occurring. The acoustic model represents the environment in which the speech takes place (over the telephone, broadcast, or microphone, etc.). Lastly, the dictionary is a list of words that can be recognized along with their pronunciation. Using different models and dictionaries specific to a particular domain can increase word recognition accuracy. ● Training tools available: The models mentioned above are generated from large sets of training data. The CMUSphinx toolkit includes several tools for generating models from scratch as well as tools for adapting existing models for a specific domain. ● Standard models available: There are several standard language models, acoustic models and dictionaries available for download. ● Portable (Java) ● Time-stamped	● Requires significant resources: Depending on the parameters used in the configuration file, this tool can require significant memory and processing power. ● Audio files must be in particular format (16 khz, 16 bit, mono, little-endian)

The configuration used for transcription has a large influence on the time required to transcribe a given file. For the configurations we found to achieve the highest accuracy, the average time required to transcribe the 8-minute audio clip was about 45 minutes. After several attempts to configure Sphinx-4 with different models and dictionaries, the lowest WER we achieved was 0.76 and the highest accuracy was 0.67. From a human perspective, this transcript was insufficient to provide a meaningful context of what took place during the meeting and was often not even readable.

Experiment 2: MAVIS, AT&T, Dragon, Google Voice, Julius

This experiment evaluated five tools: MAVIS, AT&T, Dragon Dictation, Google Voice and Julius. The metric used for these experiments was a comparison of major errors, minor errors and proper noun errors.

Major Errors:

- Continuous stream of incorrect words
- Continuous stream of missing words

Minor Errors:

- One word error
- Spelling error
- Grammar error (two/too)
- Capitalization error
- Period or thought break error
- Commas are not counted as minor errors

Proper Noun Errors:

- Inability to identify proper nouns correctly (USCB, California, Names, Senator). Proper noun errors are counted as either a part of minor or major. They are major if their context includes a major error, minor otherwise.
- We consider a noun not being capitalized an error, because Natural Language Processing software relies on correct use of nouns to identify key people and places. Therefore, we would like to minimize the number of errors that will result from Natural Language Processing software by picking a robust transcription technology.

We used a 6-minute audio clip taken from the CalChannel website. The audio file was manually transcribed in order to compare the transcripts of each tool being evaluated
ed.

MAVIS

Microsoft Audio Video Indexing Service (MAVIS) is a Windows Azure application, which uses speech recognition technology developed at Microsoft Research to enable searching of digitized spoken content. MAVIS generates automatic closed captions and keywords, which can increase accessibility of audio and video files with speech content. MAVIS uses a Deep Neural Net (DNN) based speech recognition technology, which reduces errors in speech recognition by automatically expanding its vocabulary and storing word alternatives using a technique referred to as Probabilistic Word-Lattice Indexing.

This tool is only accessible via GreenButton.com, a cloud service that hosts several applications.

Table 3. MAVIS advantages and drawbacks

Advantages	Drawbacks
● Hosted solution in the cloud ● No initial voice training required ● Better at recognizing names than other technologies ● Words that are confidently understood are in bold script ● Wide variety of input files allowed ● Time-stamped	● Not free ($15 / hour) ● Punctuation and capitalization can appear arbitrary at ● times ● Transcription of a 20 minute video can take up to 2 hours ● Words can tend to be left out altogether if not understood ● Strange characters can appear in the transcript

The resulting transcript retrieved from MAVIS contained the following errors:

- 28 major errors
- 59 minor errors
- 16 proper noun errors

AT&T

AT&Ts Speech API is a cloud-based service meant to transcribe audio to text using AT&Ts Watson speech engine. In order to do this, AT&T requires that you specify a relevant context for it to gather data from; all contexts are built into the service with no ability to specify your own context. In total, AT&T provides and maintains 7 contexts, including:

- Web search
- Business Search
- Voicemail To Text
- SMS
- Question and Answer
- TV
- Generic

Being a cloud-based service, most of the hard work is done on AT&Ts platform. As such, the API can be called from many different environments and languages to achieve the same results. Requests are made to AT&T servers through an HTTP request, which perform speech-to-text analysis on the input files using Watson speech engine. Input file formats can be of two types:

- WAV, 16-bit PCM, single channel, 8 kHz sampling

- AMR (narrowband), 12.2 kbit/s, 8 kHz sampling (recommended)

As an additional constraint, audio files can only be sent 4 minutes at a time. AT&T provides a number of APIs to use their service, supporting the following environments:

- HTML5
- MS
- RESTful

As a result, most languages can give a speech-to-text request to AT&T, including Java, Ruby, and C#.

Table 4. AT&T advantages and drawbacks

Advantages	Drawbacks
● Cheap: 1 yearly fee of $99 + $0.01 per API call past 1 million/month ● Easy to use and versatile: any language with HTTP support should be able to use it ● Works on multiple speakers ● Quick calculation: around 1 min audio / 1 min calculation	● 4 minutes at a time; must break up long text ● Not free ● Transcription is not very strong; many errors ● AMR audio format (mostly) required; W A V format worked inconsistently ● Proper noun recognition is bad: doesn't capitalize except for start of sentence, and often errors in names ● Poor punctuation: seems arbitrary at times

The resulting transcript retrieved from AT&T contained the following errors:

- 27 major errors
- 65 minor errors
- 16 proper noun errors

Dragon Dictation

Dragon Dictation is speech recognition software that lets you use your voice to create and edit text or interact with applications on your machine. It lets you use your voice to create and edit documents, manage e-mail, surf the Web, and more. It also provides digital voice software for mobile devices that let you capture your notes ”on-the-go” and transcribe them with Dragon Dictate.

The software requires configuration and depends on the user correcting its dictation as it's used. The more it is used, and the more it's corrected, the better and more accurate its language model becomes. You can even use recordings that you've made

on your mobile device in order to build your personal language model.

Although Dragon appeared to be a solid transcription technology for a single user, it proved that it was intended for exactly that: a single user. Output from Dragon also did not have any punctuation. For our purposes, it is not worth pursuing further evaluation of Dragon.

Table 5. Dragon Dictation advantages and drawbacks

Advantages	Drawbacks
● Relatively malleable language model ● Transcribes audio relatively quickly ● Can easily load audio files with a range of different	● Not free ● Requires voice training ● Intended to learn a single users speech patterns ● No punctuation ● Proper nouns may get lost in the noise

The resulting transcript retrieved from Dragon Dictation contained the following errors:

- 45 major errors
- 35 minor errors
- 16 proper noun errors

Google Voice

The Google Voice API is a speech recognition API that supports audio to text automation. It allows you to use your voice to create and edit text or interact with applications on your machine. Google Voice has its own software and also provides the framework and essence of the Closed Captioning feature on YouTube. The software is often used to translate voice mail messages to text in order to provide a message to the user without the user having to listen to it.

The Google Voice API can also be found in Android mobile phones, which it provides for Speech Recognition and navigation through applications on the phone.

This version of the Google Voice API is actually not public and can support any size videos. A Speech2Text program was written using this version of Google Voice API function calls, which takes in a WAV file and outputs the text it transcribes from the audio file. The software still has a few rough edges and also a fatal flaw when trying to process audio files with sections of little or no sound (variability in frequency).

Table 6. Google Voice advantages and drawbacks

Advantages	Drawbacks

● Transcribes audio relatively quickly ● Free ● Can transcribe any length video	● Only supports W A V files ● Has trouble with audio files that have sections of little or no sound ● No punctuation ● Proper nouns may get lost in the noise

The resulting transcript retrieved from Google Voice contained the following errors:

- 34 major errors
- 43 minor errors
- 25 proper noun errors

Voxforge / Julius

Voxforge is the most complete open-source English speech corpus; it compiles speech into acoustic models for other software systems such as: Julius, Sphinx, and HTK to work with. Using this data, these software systems can match certain sets of the resulting acoustic model to words, or perform other operations on them.

Julius is an open-source speech recognition system; its development began in 1997 in Japan and since has been refit to work for many different languages. Julius requires two things to interpret speech: an acoustic model, which Voxforge provides, and a grammar of words to match the audio against. The grammar, however, must be tailored to the acoustic model, and few generic grammars seem to exist; as such, the Julius/Voxforge combination seems like a difficult option, or one that might require more time to get setup and evaluate.

Conclusion

Within experiment 1, it is obvious that YouTube outperforms our configuration of Sphinx. Sphinx's transcript was barely readable, achieving only 0.76 for WER and 0.67 for accuracy. YouTube achieved a far lower WER of 0.29 and a better accuracy of 0.76.

Within experiment 2, MAVIS appears to be the most accurate tool with the fewest amounts of combined errors.

These results point to both YouTube and MAVIS as being the best choices for automatic transcription tools. However, we conducted one last test to gain a consistent comparison between these two tools. We ran MAVIS on the same audio file used for YouTube, and calculated WER and accuracy scores. MAVIS achieved a slightly better WER of 0.27 and the same accuracy of 0.76. MAVIS appears to be the strongest choice.