1. Introduction: CalMatters and Digital Democracy Project

“CalMatters is a nonpartisan and nonprofit news organization bringing Californians stories that probe, explain and explore solutions to quality of life issues while holding our leaders accountable. We are the only journalism outlet dedicated to covering America's biggest state, 39 million Californians and the world's fifth largest economy.”

2. Digital Democracy Website and Database

The Digital Democracy project was inspired by the decline in media coverage of California's state government and by the significant impact of decisions made in a state that ranks as the world's fifth largest economy with an annual budget of about $300 billion. The Digital Democracy website is a portal for a general, non-expert audience to understand the people and activities in state government so there is a better opportunity for civic engagement and a healthier environment for decisions in the public interest. It is also intended to help journalism by providing reporters with quick and easy access to a vast amount of information.

Digital Democracy was built in 2023 in a collaboration that included CalMatters, Cal Poly San Luis Obispo and 10up, a global web development firm. The database and artificial intelligence were designed and built by faculty and students at Cal Poly, where an earlier version of Digital Democracy was launched in 2015 by former state Sen. Sam Blakeslee and then Lt. Gov. Gavin Newsom at the university's Institute for Advanced Technology and Public Policy. In 2023 the project was moved to CalMatters.

The Digital Democracy website introduces each of the 120 California legislators and displays the current year's agenda on six major topics, as seen in legislation. Our technology tools collect data throughout state government about legislation, votes, money, and politics as well as the transcripts of public hearings and floor sessions. That unique database is searchable from our website.
Finally, to help journalists identify stories about the legislative process, we built a custom artificial intelligence tool that scans the database to suggest story ideas for reporters. When reporters pursue these tips, the resulting stories appear on the web pages for legislators as well as on CalMatters. They are also shared with media throughout the state, as are the story ideas for reporters identified by the AI tool.

All information regarding hearing, bills and people is collected from various sources. The majority of the information comes from official state websites. All of this data is stored within a central database for the Digital Democracy Project (known as DDDB). This database currently has over 200 tables storing not only records but also the relationships between the disparate records collected. In addition, many of the tables are ‘derived tables’ which store aggregated data in a manner which can be easily and quickly retrieved by the website.

The course project for this class will focus on the data contained within one particular source, namely the CalAccess Database maintained by the Office of the California Secretary of State.

3. CalAccess Database

Part of the Digital Democracy Database is constructed out of the information publicly shared by the office of California Secretary of State. This information is accessible from a database called CalAccess (which we sometimes also refer to as the "SoS Database"). CalAccess provides public access to its data via a collection of TSV files (which you will eventually parse for the data to be stored in the LEAKS DB). Much of same information can be also viewed via the CalAccess Web Portal.

CalAccess database is based on the content of a variety of forms that organizations and individuals involved in political activity in California must file with the Secretary of State office. Each filer is given a filer ID (Some filers may have 2 filer IDs since the SoS Database has changed over the decades). Most of the data provided by the CalAccess database is provided as a digital representation of a submitted form. Each form has a Filing ID and an Amendment ID in order to correctly identify the current information for a filer.

CalMatters and the Digital Democracy project are interested in collecting the following types of information from the CalAccess database:

- **Information about Registered Lobbyists and Lobbying activities.** State of California defines Lobbyist as
  
  "... an individual who is compensated to communicate directly with any state, legislative or agency official to influence legislative or administrative action on behalf of his or her employer or client."

1 https://www.fppc.ca.gov/learn/lobbyist-rules.html
California requires all individuals who spend more than ⅓ of their time doing advocacy as described above to register with the state. The Secretary of State office also requires, through a variety of forms, disclosures about their employment, and their lobbying activities in the California Legislature (i.e., the Bills the Lobbyists are paid try to get passed, amended, or blocked).

- **Information about Candidates running for elections in the State of California.** Candidates running for Federal, State, and Local elections in the State of California must file paperwork with the office of the Secretary of State (this is how they get on their respective ballots). The Digital Democracy project is interested in retrieving this information from the CalAccess Database.

- **Information about Financial transactions associated with Elections in the State of California.** As part of efforts to ensure transparency of elections, the State of California requires all campaign committees (organizations formed to support or oppose the election of a specific individual, a ballot measure, or advance specific policy goals - see notes about different types of committees in Section 5) and sometimes - other organizations - must disclose all financial transactions associated with their election activities. These include the contributions to the campaign committees from different organizations and individuals (this is how election campaigns are funded), independent expenditures that outside organizations undertake targeting specific candidates, and campaign committee expenditures that the campaign committees undertake (this is how they spend the campaign contributions). CalAccess database stores the information about these disclosures.

**Need for Project CalAccess LEAKS.**

In this project you will be fulfilling two complementary roles: (a) a second pair of eyes for some of the information contained in the CalAccess database, that is already imported into DDDB, and (b) developers of prototype software for incorporating into the DDDB the information that is not there, or that is being ingested through alternative means. We explain these two roles below.

**A second pair of eyes. First,** the Digital Democracy Database already contains some of the information about Lobbyists, Elected Candidates and Election Financials, but this information is not always collected in the most efficient, or convenient way. Your solutions will provide a "second pair of eyes" on some of the design decisions, and will, when integrated with the Digital Democracy Database, improve the overall performance of the software built as part of the Digital Democracy project and the health of the Digital Democracy Database.

**Developing prototype software. Second,** some of the information that the CalAccess database contains and we want has not yet been incorporated into the Digital Democracy Database, or (see below) has been incorporated through another intermediary source, which we want to stop using in the future. Your project(s) will enable us to enhance the content of the
Digital Democracy Database, and provide additional support of the Digital Democracy family of software applications.

More specifically, currently Digital Democracy receives election financial information from a third party organization “Open Secrets”. This organization performs some tagging and categorization which is useful to us. First they associate campaign committees to candidates within elections. Second they associate Organizations with industry category groups. This categorization is what powers the “sunburst” chart on the Digital Democracy legislator pages.

However, these benefits come with certain drawbacks. The primary drawback is the time delay between when CalAccess publishes data and when Open Secrets is finished processing/ingesting the data (up to a matter of months). This delay is not ideal for CalMatters as a news organization, which strives to write timely stories. The second drawback is a lack of data transparency. Currently Open Secrets does not provide publicly an easy way to link financial data back to the source filings. This makes it more difficult for CalMatters to fully cite exactly where the data in their stories originated from.

4. Project CalAccess LEAKS

For your project, you will complete four major tasks. Two of the tasks are database design and implementation, and two - software design and implementation (basically - you build a database, and write scripts that bring the data into the database).

These sets of tasks are outlined below.

**Task 1.** Design and implementation of the **Primary LEAKS Database.** The "Primary LEAKS Database" is a database that will capture all the information from the CalAccess Database related to Lobbyists and their activities, election candidates for California elections, and financial transactions associated with election activity.

**Task 2.** Design and implementation of software (a collection of scripts) that parses the information for the CalAccess database (either from CalAccess TSV files, or from a MySQL database where each table is essentially one CalAccess TSV file). Taken together these scripts shall capture all the information we need from the CalAccess database, and populate, and regularly update the tables of the Primary LEAKS Database. (Note: you may get some assistance with setting up the "regularly update" part of your software).

**Task 3.** **LEAKS Data Warehouse.** The CalMatters Digital Democracy web application, and the CalMatters TipSheets web application will be using aggregated information that can be derived from the data you place into the Primary LEAKS Database (for example, given a Bill, the application will want a summary of all lobbying activity surrounding it - e.g. how many lobbying
contracts reported to the office of the Secretary of State mention the Bill). In the second part of your project, you will design and develop a LEAKS Data Warehouse Database, designed for read-only access to the summative data.

**Task 4.** Design and implementation of software (a collection of scripts) populating the tables in the LEAKS Data Warehouse based on the data stored in the Primary LEAKS Database.

This document concentrates on providing you the requirements and the specifications for Task 1 and Task 2. In a few weeks (after the design of your Primary LEAKS Database is complete), we will release documentation specifying requirements for Tasks 3 and 4.

---

5. CalAccess Data that Needs to be Captured in the Primary LEAKS DB.

As stated above, there are three major types of data we are interested in capturing from the CalAccess Database: Lobbying Activity, Election Activity/Candidate Filings, and Financial Activity associated with Elections. Below, we provide some information of what needs to be captured for each part of the database.

5.0. Information available in the Digital Democracy Database

Some of the information that you will be collecting from the CalAccess database will involve references to outside entities that *by themselves are not tracked in the CalAccess DB*, but that exist in the Digital Democracy Database. We will expose for your using in the project the Entity Sets (and later - the DDDB tables, or their simplified versions) that are referenced in such a way. A preliminary list of such Entity Sets is provided below (Note: we may discover that we need to expose additional Entity Sets to you at some future point - in which case additional documentation will be released). For now, you can assume that each Entity Set mentioned in this Section has a unique primary key attribute. Additional attributes for these Entity Sets will come from your observations of how they are referenced in the CalAccess database.

**Bills.** Some forms that need to be filed with the Secretary of State Office contain information about legislative activity that will be lobbied. These forms reference bills that are being
considered in the California Legislature. The forms will use a "XXX NNNN" format when referencing the bills, where "XXX" is a bill prefix, and "NNNN" is the bill number. Bill prefixes specify circumstances under which the Bills originate: e.g., "AB" stands for "Assembly Bill" (i.e., a regular Bill originated in the California State Assembly), while "SB" stands for "Senate Bill" (i.e., a regular Bill originated in the California State Senate). "NNNN" is the Bill number. The full reference will look like "SB 23" or "AB 145". The full list of Bill Prefixes exists in the DDDB and can be made available to you (it might make parsing a bit easier). Bill Numbers as specified above, however, are NOT unique. California Legislature has two-year legislative sessions after each state elections, and in each legislative session, the bill numbering starts anew. So, "SB 1" during the 2023-2024 session is a different bill than "SB 1" during the 2021-2022 session. Legislative sessions last two years and always start on an odd year. Any forms filed with the Office of Secretary of State that mention a Bill Number always refer to the Bills in the current session. In a vast majority of cases, checking the year of the filing (or the full date) will inform you on the current session. Any discovered edge cases will be discussed and documented separately. The DDDB database, in addition to the unique Bill Id (DDDB Bill Ids are unique), has additional information about the bill and its multiple possible versions, including all sorts of textual descriptions of different Bill versions (capturing the legislative process).

Lawmakers. Winning election candidates in State Legislative Elections become Lawmakers. The DDDB associates each Lawmaker with a specific Term in the office during which they maintain the Chamber (Senate or Assembly), their District (e.g, AD-20, or SD-17), Party affiliation (any mid-session party changes are handled separately along the way, you do not need to worry about it). For election candidates who are incumbent lawmakers, and who won elections and became lawmakers mentioned in any of the filings (candidate filings, different financial transaction filings by their campaigns), we always want to associate all information available in the CalAccess database with their identity in the DDDB. Lawmakers have names (first, middle, last), and this is how they are often referred to in the CalAccess filings whenever they are referenced.

Lobbyists. DDDB maintains a Lobbyists entity set, which assigns each Lobbyist a unique internal DDDB Person Id (same type of Id as assigned to each Lawmakers). All other information about a Lobbyist comes directly from the CalAccess database.

Persons. Lawmakers and Lobbyists are two kinds of people the Digital Democracy Database tracks. In general, DDDB maintains a large list of people who participated in the legislative process. In addition to the Lawmakers and Lobbyists, the Persons entity set contains information about various individuals who testified in the legislative hearings. At the same time,

---

2 There is a short period of time after the elections (November - December of an even year) when some Bill Numbers for the next year's session may already be introduced. If we/you discover that this is an issue with the Secretary of State filings, we will issue more detailed instructions on how to handle it. For now simply checking the year of the filing will inform you of the true identity of each bill.
many different types of filings with the Secretary of State office contain mentions of individuals who are neither Lobbyists, nor current or past Lawmakers. The two largest sources of individuals beyond Lobbyists and Lawmakers are Election Candidates who lost elections, and individuals who donated money to different campaigns. For these two types of individuals, it is highly desirable to determine if their records are in DDDB, but this determination is out of scope of your project.

Organizations. Like individuals, Organizations (companies, municipal/county governments, associations, unions, etc…) actively participate in political activities, and where their participation crossed the "view frustum" of DDDB, the Digital Democracy project collected the information about them. The DDDB database only maintains a unique internal ID of each organization and a set of names under which the organization is known/has been referred to (e.g., the DDDB will store both "ACLU" and "American Civil Liberties Union" and associate both with the same organization, although full mapping of mentions to organizations is currently work in progress). in CalAccess, organizations appear in a wide range of filings: as sources of campaign funding, as Lobbying Firms or as direct employers of Lobbyists (see below). CalAccess is the authoritative source of information on Lobbying Firms. All other Organization mentions are considered references to Organizations in the DDDB. Full and unequivocal matching of all organizations referenced in the CalAccess database to their DDDB counterparts is out of scope for this project, but we can and will provide you with some lookup functionality that can resolve many of the Organization mentions.

5.1. Lobbying Activity

Lobbyists. All persons and firms engaged in what the State of California defines as official lobbying activities must be registered via filing of a form (Forms 601 and 604) with the office of the Secretary of State. Each Lobbyist and lobbying firm is assigned a unique identification number (often referred to as a Filer ID). Newer FilerIDs are fully numeric, however, in the past, Filer IDs were combinations of letters and numbers. Anywhere, where we mention FilerIDs in this document we potentially mean a pair of FilerID numbers - an old, and a new one. The CalAccess Database contains full content of each filed form 604 (Lobbyist Certification Statement). These filings contain the first, middle, and last name of each Lobbyist (the also might contain other information related to an individual's name, such as their suffix such as "Jr" or "III"). It also records whether or not the lobbyist has completed the required lobbying ethics course.

Lobbyist Registration. Lobbying Firms and Lobbyist Employers must file a registration form for each time period (typically between November 1st and December 30th every even numbered year) during which they want to be able to conduct lobbying activities. This form is often attached to a lobbying firm’s registration form (Form 601) or a lobbying employer’s registration form (Form 603). Each form provides Lobbyist registration information for a specific time period. The Primary LEAKS DB shall capture information about the years when a Lobbyist was
active/registered. If a Lobbyist had a registration gap, or stopped their registration altogether, these facts must be observable from the data stored in the Primary LEAKS database.

**Lobbyist Employment Records.** Lobbyists are employed in two different ways. Some Lobbyists are permanent employees of the so called Lobbying Firms - organizations whose main purpose is provision of lobbying activities for third parties. If a Lobbyist is employed by a Lobbying Firm, the Lobbying firm files a disclosure (Form 601) of the fact of employment. These are typically filed on an biannual basis for Lobbyists with continued employment, or within a certain period of time mandated by law, when a Lobbyist changes their place of employment. Another way for a Lobbyist to be employed is direct employment (usually in the office of Public Affairs, or Public Policy) by an organization that desires to engage in lobbying activities. In this case, we say that a registered Lobbyist is *directly employed* by the organization they lobby for. To explain this a bit better, here are two simple scenarios of how lobbying employment works. Direct Lobbyist Employment is reported (Form 603). For both employment by the Lobbying Firm as well as Direct Employment, we need to know the employment dates associated with the filings.

**Example Scenario 1.** Lobbyist Bob Smith is employed for a Lobbying Firm Smith, Brown and Tanner. This firm specializes in representing municipal governments in Sacramento lobbying for their positions on bills that relate to separation between State and Municipal regulations. San Luis Obispo County government can hire Smith, Brown and Tanner to represent their position on a Bill regulating the use of off-road vehicles in State Recreational Parks (SLO County Government has a vested interest in this bill due to the presence of the Oceano Dunes State Park within the county boundaries).

**Example Scenario 2.** Lobbyist Helen McDonald is hired to be part of the Public Relations department of General Electric. GE is interested in lobbying on multiple bills regarding the energy production in California, and perceives the amount of lobbying activity to be sufficiently high to have a permanent employee representing their interests, rather than to hire a Lobbying Firm. In such a case, we say that Helen McDonald is a Lobbyist who is directly employed by GE.

**Lobbying Firms.** Like any entity in the State of California that is required to file forms with the Office of the Secretary of State, Lobbying Firms have individual *Filer ID* numbers. They also have names and addresses.

**Organizations (Direct Employers).** Pretty much any organization that is NOT a Lobbying Firm can directly employ a Lobbyist. Such organizations also have individual *Filer ID* numbers and names.

**Lobbying Activity.** Any organization that hires a Lobbying Firm to lobby for it, must file a Lobbying Activity disclosure form (Form 603). This disclosure form has information about the date the contract started and the duration of the contract. In addition, Organizations file a
quarterly form (Form 635) discloses all Bills and government agencies lobbied that the organization has initiated lobbying on as part of their representation of the organization that contracted it. Note that there may be many Bills listed in a single contract (when you see the data, one of the technical problems for you might be proper parsing of all bill numbers).

We also note something that is not available in the CalAccess database. Whenever an organization hires a Lobbying Firm to represent it, the contract (and the appropriate form disclosing it), establishes a relationship between these two entities. No filings disclosing the names of specific Lobbyists who work for the Lobbying Firms and will be lobbying on behalf of the organization that contracted the firm are required. As a result, the CalAccess database does not allow us to associate individual Lobbyists with a specific contract or a specific Bill on which they might be lobbying. (The DDDB gets this information from observing the appearances of the Lobbyists in legislative hearings, but this is not part of the scope of your work).

Finally, some entities do not hire firms or lobbyists but still send money advocating for or against legislative proceedings. If these entities spend more than $5000 in one quarter they must also file with the Secretary of State (Form 645).

5.2 Election Candidates

The CalAccess database maintains information about all elections in the State of California. If an election appears on any ballot in the State of California, the Office of the Secretary of State is responsible for administering it - registering the candidates, preparing the ballots, etc.

Elections. We concentrate on State and local elections here (leaving elections of the President, and the US Congress out of the scope of this project). In California, the major types of elections are:

- **Primary:** This election happens before general elections and serves to lower the amount of candidate choices in the general.
- **General:** This election is the normal way in which a candidate is elected to office usually occurring in November in even-numbered years.
- **Recall:** Under certain circumstances an elected official may be the subject of a recall election. In this case the voters choose yes or no on the ballot whether or not the official should be removed from office.
- **Special Election:** A non-standard election in which a candidate is chosen for office. Special elections can happen for a variety of reasons.
- **Runoff:** An election which occurs after one of the previous elections which is triggered under certain conditions to decide again the outcome.
Elections for some offices are two-stage - first a primary election is conducted to determine the two candidates who advance into the final election round.

In addition to its type, each Election has a date it takes place on (or at least - a year). Thus, we can tell apart the Governor's elections in 2020 from the Governor's elections in 2024.

Offices. Many of the elections above are associated with a specific office - either a state-wide elected executive office, or a legislative office, or a municipal or judicial office. An Election for an office is different from the office itself in that an Election is a specific event (with a date), and given an office, over a period of time, there may be multiple elections for it. Note that recall elections are also associated with an office.

Candidates. For most elections, the Candidates are individuals who want to be elected to the office (or individuals who want to remain in the office). Recall Elections and Propositions don't have candidates per se, instead they have positions which can be supported, like Candidates in other elections by various groups who will spend money during the campaign. Like all other entities that file forms with the Office of the Secretary of State, Candidates have their Filer ID, first, middle and last names.

Controlled Committees. To spend money on an election, a Candidate must establish a Campaign Committee (which they control). The election expenditures are made by the committees. Committees have a name (such as "Bob Smith for Senate 2024"), and are associated with a Filer ID and an Amendment ID.

Out of scope. While the Office of the Secretary of State conducts and certifies elections, and the voting information is available from the Office of the Secretary of State, for this project, tracking the results of elections is out of scope.

5.3. Election Financials

The Office of the Secretary of State maintains information about financial transactions that are made in support of electoral campaigns of candidates and positions (for Ballot Initiative/Proposition elections). Financial transactions we are interested in capturing come in three different "flavors":

- **Campaign Contributions.** These transactions document the money a Donor gives to a Campaign Committee of a candidate or a Campaign Committee advocating on behalf of a Proposition/Ballot Measure.
- **Expenditures.** These transactions document the money a Campaign Committee spends on election-related activities.
• **Independent Expenditures.** In these transactions, **Donors** spend money directly on behalf of a Candidate or a Proposition/Ballot Measure, without contributing the money to a Campaign Committee.

**Examples.**

- California Chamber of Commerce donates $2000 to the "Bob Smith for Attorney General" Campaign Committee. This is a campaign contribution.
- "Bob Smith for Attorney General" Campaign Committee spends $40,000 to purchase TV advertising. This is a campaign expenditure.
- California Chamber of Commerce spends $20,000 to print and send out mailers advocating in favor of CA Proposition 34. This is an independent expenditure.

Please note that two of these three types of transactions - namely, campaign contributions and independent expenditures are spending targeting the campaign (independent expenditures is un-coordinated with the campaigns spending for the benefit of the campaign), while the remaining type - campaign expenditures - represent the outflow of the money (expenditures are made with the money that came to the campaign committee as a campaign contribution).

As seen above, each type of financial transaction has a Spender - the entity that "parts" with the money. Campaign Contributions are associated with the target Campaign Committee that receives the money. Both types of expenditures have "target purposes" (what the money was spent on. Below we provide a more detailed overview of each type of financial transaction and on Campaign Committees.

**Campaign Committees.** State of California recognizes four types of Campaign Committees:

- **Committees controlled by the Candidate.** These committees are formed by the Candidates who run for elections. The Candidates and their surrogates control directly the spending of the funds by the Committee. These committees are explicitly associated with a Candidate standing for a specific election. Candidates can control multiple such Committees.
- **Candidate or officeholder primarily formed committee:** these are committees formed by "third parties" not directly affiliated with the Candidate/Officeholder, whose primary purpose is to support or oppose a single candidate in an upcoming election.
- **Ballot Measure Committees.** These Committees are formed to advocate in favor or in opposition to a specific Ballot Measure. These Committees are associated with a specific Ballot Measure/Proposition. If possible/available, it might be useful to associate their position (support/opposition) on the specific measure/proposition if it is available.
- **General Purpose Committees.** These are the Committees that spend money on political activities that are not tied to a specific candidate.

For example, California Teachers Association, may want to advocate for additional benefits for the California K-12 educators, and support multiple candidates in the elections who share its positions (and also - oppose the candidates who do not share
those positions). In such case, CTA would form a general purpose committee, and direct this committee (with the funds received) to make independent expenditures on behalf of a number of different candidates.

Campaign Committees of different types must file different disclosure forms with the office of the Secretary of State (Forms, 400, 410) whose contents are available in CalAccess. Information about the Committees found in those forms includes their name and address. Additionally, we want to keep track of the date the initial disclosure form for each Committee was filed.

**Campaign Contributions.** Campaign contributions are reported to the SoS Office on Form 460 by the controlled Committees that receive the Contributions. Most contributions are monetary and are reported in Schedule A of the form (Form 460 has many different parts). Each Campaign contribution is associated with a Donor (Giver). Givers can be individuals or organizations. Individual givers can be classified as major donors (defined as individuals making more than $10,000/year in total monetary or non-monetary contributions), general public (individuals donating less than $10,000/year), or the candidate themselves. Similarly, organizations also are divided into major donors and "other" using the same criterion as individuals.

**Givers.** For individual givers, the form records information about their name (First, Middle, Last), their address (State, City, Zip), and their Employer, which is an Organization. For Organization Givers, there are four types we can establish from the form in addition to the information above:

- Controlled Committee (i.e., the Controlled Committee for one Candidate donates to the Controlled Committee for another Candidate - e.g., the Controlled Committee supporting a candidate for the Governor, donates to the controlled committee supporting the Governor's preferred choice of Lieutenant Governor).
- Party Committee (e.g., State Democratic Party donating to a Democratic candidate for Assembly District 24)
- Independent Expenditure Committee ("Californians against Prop 25” donating to a candidate running for the position of the Insurance Commissioner)
- Other Organization (any non-Committee organization, e.g., Chamber of Commerce donating to a candidate in Senate District 4).

**Expenditures.** Expenditures by Controlled Committees are disclosed in Schedule E of Form 460. The form is filed by a Controlled Committee, so the expenditure is associated with the Controlled Committee that filed the form. We want to record the amount of money spent, the date of the transaction (as opposed to the date the form itself was filed), and the information about the Receiver of the expenditure and the purpose of the expenditure. The receiver of the expenditure is either an Organization (e.g., the Committee paid a marketing firm for campaign mailings or text messages) or an individual (the Committee paid salary to its employees). The
type of expenditure is controlled by an Expenditure Code (the list of possible codes is fairly large, available on the form 460, and will be shared with you - a code is typically a three-letter combination like "WEB" or "LEG", or "PHO"). Notice that not every expenditure will have a code associated with it (some Committees file expenditure reports where some expenditures don't have codes).

**Independent Expenditures.** Independent expenditures are reported to the SoS Office in Form 461. These expenditures are associated with:

- the Campaign Committee (not controlled by the Candidate on whose behalf the expenditure is made) that made the expenditure
- the Receiver of the expenditure (similarly to Campaign Expenditures - an individual or an organization that received the money)
- the Target - a Candidate, or Ballot Initiative Position on whose behalf the expenditure is made.

An example of an independent expenditure is a "Californians Against Prop 4" General Purpose Committee paying $20,000 to "ABC Marketing" for completing a series of phone calls in Assembly District 24 on behalf of candidate Emily Rogers running for the AD-24 seat in the upcoming election.

Independent expenditures, similarly to Campaign Expenditures come with a code identifying their type (and implicitly - whether the Reciever is an individual or an organization; as with Campaign Expenditures, the codes may not be present for some expenditures), transaction date and transaction amount, as well as the purpose of the transaction.

### 5.4. Additional Notes

We plan to share examples of different forms mentioned in this document with you on or around Thursday, April 18. The forms may contain additional attributes for different entity sets and relationship sets that are not directly mentioned in this document. As a result, we acknowledge that while this document provides sufficient information to identify all entity and relationship sets that need to be represented in the Primary LEAKS DB, specific details about some of them may evolve over time.

### 6. Task 2: LEAKS DB Data Ingestion Scripts

The CalAccess database consists of multiple (in fact - numerous) spreadsheets, stored and made publicly available as Tab-Separated Values (TSV) files. Most spreadsheets available from the CalAccess database document filings associated with a specific form, but some spreadsheets combine information from multiple forms into a single table. Different rows in such spreadsheets may represent different forms filed. Such forms would usually share a significant number of fields (columns in the spreadsheet).
In addition to this, the Digital Democracy project maintains an up-to-date version of the CalAccess database stored in MySQL. The Digital Democracy version of CalAccess database creates a single relational table for each CalAccess TSV file. All columns are VARCHAR and are not parsed. This is done for ease of access to this data.

You will be provided with access to both the spreadsheets (they are publicly available) and the Digital Democracy CalAccess DB clone.

The key objective of the LEAKS DB project is for you to design a database that will hold all the information specified in Section 5 of this document, and to write software that extracts this information from the CalAccess database (either the TSV files directly, or from the Digital Democracy CalAccess MySQL clone), and places it into the Primary LEAKS DB.

The software part of this project is a collection of scripts (we strongly recommend Python, or a combination of SQL and Python, if a team has reasons to use another programming language, please consult us first) that populate and maintain the LEAKS DB database.

In a stable state, your software shall access the CalAccess source, determine if there is new information added to the CalAccess database and export this information into the Primary LEAKS DB. When run as part of the Digital Democracy codebase, these scripts will become part of a regular (usually, nightly) DDDB update cycle. Currently, this is done using cron. In the future the Digital Democracy project may switch to Github Actions. We will provide some additional guidance for how to set up the scripts to be runnable in a "nightly update" fashion.

7. Tasks 3 and 4: Data Warehouse and Data Warehouse Creation Scripts

The LEAKS Data Warehouse shall contain tables that provide Digital Democracy API with immediate access to the desired information via simple SELECT-FROM-WHERE queries that involve only a single table and only some filters in the WHERE claus. The specific design of the Data Warehouse tables depends on the specific Use Cases each team will be assigned in Lab 4. At that time, each team will receive its own list of Use Cases outlining the information needs, and will design the appropriate Data Warehouse E-R model for their part of the warehouse.