SPENDVIEW: A platform for democratizing access to government budget and expenditure data

Manuel Aristarán

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Master of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology

June 2016
© Massachusetts Institute of Technology, 2016. All rights reserved.

Signature redacted

Author
Manuel Aristarán
Program in Media Arts and Sciences
May 5, 2016

Signature redacted

Certified By
Dr. César A. Hidalgo
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Signature redacted

Accepted By
Dr. Pattie Maes
Academic Head
Program in Media Arts and Sciences
SPENDVIEW: A platform for democratizing access to government budget and expenditure data

Manuel Aristarán

Submitted to the
Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Master of Science in Media Arts and Sciences
at the Massachusetts Institute of Technology
June 2016

Abstract

Budgeting and expenditure data is the clearest expression of a government's priorities. Despite its importance, making it available to the public imposes hard challenges that not every administration is ready to undertake. The lack of IT capabilities and constrained resources of government agencies —regardless of size or budget— make it difficult for them to respond to the demands of information from their constituencies, transparency advocates, the press and central governments. Moreover, these administrations don't usually have access to analysis tools that help them view how the resources are being used and to detect potential risks of misspending —a critical need of elected officials, who are under everyday scrutiny. Responding to these needs, I propose the design, implementation, impact and usability studies of a software platform for analysis and visualization of government budget and expenditure data.

Thesis Supervisor

Dr. César A. Hidalgo
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences
SPENDVIEW: A platform for democratizing access to government budget and expenditure data

Manuel Aristarán

Signature redacted

Thesis Supervisor

Dr. César A. Hidalgo
Associate Professor of Media Arts and Sciences
Massachusetts Institute of Technology
SPENDVIEW: A platform for democratizing access to government budget and expenditure data

Manuel Aristarán

Thesis Reader  
Signature redacted

Ethan Zuckerman  
Principal Research Scientist — Center for Civic Media  
Massachusetts Institute of Technology
SPENDVIEW: A platform for democratizing access to government budget and expenditure data

Manuel Aristarán

Signature redacted

Thesis Reader

Dr. Iyad Rahwan
Associate Professor of Media Arts and Sciences
Massachusetts Institute of Technology
Acknowledgments

This thesis is dedicated to my wife Luisina Pozzo Ardizzi, partner in life and accomplice in countless crazy adventures. None of this would have been possible without your love and support.

When I first got connected to the internet, the coolest and craziest things were hosted in the mit.edu domain. It was through listserv archives coming from the mythical 18.0.0.0/8 class-A IP network that I was first exposed to the hacker culture. Sneakily typing commands on a VT-100 terminal in the computer lab of the university of my hometown in southern Argentina, my 14-year old self fantasized about this mysterious and faraway university in the USA.

Twenty years later, I met César Hidalgo during a fortuitous visit to the MIT Media Laboratory, and I asked him what I needed to do to study at his group.

—“Application period opens in September, huevón.”, he answered.

I didn’t think that I had a chance to study here: it was MIT, home to Minsky, Papert, Shannon and Chomsky. The word hacker was coined here! Besides, I was 33. But my wife talked me into trying (she can talk me into doing anything). I took the English sufficiency test, painstakingly crafted the admission essay, put the paperwork together and submitted my application. Two years after, I’m about to graduate from this magical place, but still find it hard to believe that César trusted me and accepted me as a student of Macro Connections. I will always be grateful for his confidence in me, and for always encouraging all his students to think big, work hard and go farther.
Macro Connections gang, you will always be in my heart. Cristian Jara Figueroa, Kevin Hu, Dominik Hartmann, Flavio Pinheiro: thank you for the laughs, your friendship, your support, your always insightful feedback and kind criticism, and the infinite number of ping-pong matches. I’m grateful for the friendship of the rest of my academic siblings: Mia Petkova, Amy Yu, Mary Kaltenberg, Elisa Castañer, Ambika Krishnamachar, Miguel Guevara and Jermain Kaminski.

Thanks to my fellow Media Lab students, faculty and staff. You have truly made me feel at home away from home. Special acknowledgments to my thesis readers, Ethan Zuckerman and Dr. Iyad Rahwan for their insightful feedback and encouragement during the process of writing this thesis.

I would also like to acknowledge the government of Bahía Blanca, my hometown, that through its Innovation and Open Government Agency allowed me to test the ideas, concepts and tools outlined in this thesis, using the city’s data. Uruguay’s Budget and Planning Office and the Latin American Initiative for Open Data for hosting me and facilitating my research in the beautiful Montevideo.

Finally, I would like to thank my family: my father Roberto, my mother María Inés, my brother Agustín and my wife Luisina. Gracias for always reminding me that love is the only thing that matters.
Table of Contents

Introduction ............................................................................................................. 15
Motivation .................................................................................................................. 16
The nature of fiscal information .............................................................................. 18
Contribution ............................................................................................................. 19

Background ............................................................................................................. 21
Prior Work ............................................................................................................... 22
The Relevance of Dimensional Modeling .............................................................. 27
Multidimensional Modeling of Fiscal Information ................................................. 30

Implementation ...................................................................................................... 35
Design Rationale ..................................................................................................... 36
  Design principles .................................................................................................. 36
  Narratives ............................................................................................................. 36
  Hypermedia ......................................................................................................... 36
  Internationalization ............................................................................................ 37
  Openness and reusability .................................................................................... 37
Mondrian-REST: a Web-friendly API for OLAP .................................................... 37
  Multidimensional eXpressions (MDX) ................................................................. 37
  A REST endpoint for OLAP databases ............................................................. 39
  Implementation Details ....................................................................................... 44
User experience and user interface ....................................................................... 45
  Overview of a Budget .......................................................................................... 45
  Drilling down on a budget classification ............................................................ 48
  Exploring the different perspectives of the budget ............................................ 49
    Budget dimensions as questions ...................................................................... 49
    Data verbalization as a complement of visualizations ....................................... 51
    Multiple graphical representations ................................................................... 51
    Exploring the data ............................................................................................ 52
    Sharing ............................................................................................................. 54
  Implementation of the user interface ................................................................. 55
    Isomorphic Web applications ........................................................................... 56
    Visibility to non-rich internet clients ............................................................... 57

Impact ...................................................................................................................... 59
Bahia Blanca: from confrontation to collaboration .................................................. 60
Working with Uruguay's budget office .................................................................. 63
  Workshop .............................................................................................................. 64
  User evaluation of SpendView ............................................................................. 65
Public Deployment .................................................................................................. 66
  Usage data ......................................................................................................... 68

Conclusion .............................................................................................................. 71
Future Work .................................................................................................................. 72
Concluding Remarks ....................................................................................................... 72
List of Figures .................................................................................................................. 73
Index of Tables ............................................................................................................... 75
Bibliography .................................................................................................................... 76
Introduction
Motivation

Government, and the public sphere in general, are active participants of the data explosion that has been happening during the last decade. In the last 10 years, numerous public entities around the world have been active participants of the Open Data movement, that seeks to publish government generated, machine readable data under licenses that allow for unrestricted use, redistribution, modification and propagation (Open Knowledge Foundation, 2015).

However, the increased availability of information resources hasn't always been accompanied by interactive tools that present this data in actionable formats, that allow end users to make informed decisions. The traditional design paradigm for the existing tools is the dashboard, that — just like those in vehicles — allow an analyst or operator to obtain information at a glance from graphical representations of the information generated by the system (Figure 1).

![Figure 1: A traditional dashboard for government finances](image-url)
Dashboards predate the Web, so they don't usually incorporate some of the interaction patterns and conventions that became predominant in the last 20 years. As they allow users to filter the information being displayed, it is usually cumbersome or impossible to share a particular state (view) of the charts and data tables. Also, as dashboards are designed to provide information \textit{at a glance}, they tend to underutilize the scroll interaction, now prevalent in mobile interfaces. Lastly, besides chart captions, typical dashboards don't incorporate \textit{text} as a way of describing data to their users.

Lately, there has been an emergence of visualization engines that improve on the classic \textit{dashboard} design logic, by adopting elements of linear storytelling and now conventional patterns of user interaction. The \textit{profile page}, for instance, is one of such conventions. In the context of a Web page, it contains information about an entity of the system (a user in a social network, an article of an online encyclopedia) and presents its information hierarchically.

Notable examples of these ideas include the latest version of the Web site \textit{The Observatory of Economic Complexity} (Simoes Gaspar, 2012), which main entities are profile pages for countries and products, and \textit{Data USA} (Datawheel, Macro Connections / MIT Media Lab, \& Deloitte, 2016) that uses the same technique for offering information about geographical entities, industries and academic majors.

Fiscal information, defined as detailed information records about budget, spending and revenue, are often recorded in multidimensional datasets, a structure that has been traditionally visualized with dashboards. This work will attempt to show that incorporating design elements such as profile pages for the entities defined in a budget, linear narratives instead of configurable visualizations, automatically generated descriptive text along with charts, indexable and shareable pages, can improve on the comprehension, accessibility and diffusion of fiscal information.
The nature of fiscal information

The data generated during the process of formulation and execution of a government's budget is one of the information resources most relevant to the public interest. An administration's policies, encoded in its fiscal measures, influence the lives of its constituents, affecting services such as health, education and utilities. Budget information can be interpreted as what is known in the economic literature as the revealed preferences (Samuelson, 1948) of a consumer, in this case a government administration. It can be argued, then, that public expenditure and budgetary records, when analyzed and visualized with effective techniques and devices, can inform the people, researchers, and public servants about the priorities of their government.

However, fiscal information is inherently complex. The budget design process is a multi-stakeholder procedure, in which disparate actors have different and often competing interests. The politician sees it as an event conducted in the political arena for political advantage, the economist as a question of resource allocation, the accountant focuses on the auditing perspective, while the public servant sees the budget as a blueprint for the implementation of policies (Lynch & others, 1979). This multiplicity of perspectives is directly translated into the structure of the information generated by the budgeting and expenditure processes: the data is recorded in mutually independent budget classifications (Jacobs, Hélis, & Bouley, 2009), each of those serving the needs of different users of the information.

The complexity of fiscal information lies on its multi-dimensional nature. This presents an opportunity to the designer of a tool for presenting it effectively; it is through the combination of these different perspectives (i.e., dimensions) that useful knowledge emerges.
**Contribution**

In this thesis I design and implement *Spendview*, a general tool to visualize and analyze government spending data.

To create *Spendview*, I developed Mondrian-REST, which is a server side software component that allows creation of HTTP APIs for accessing a database by specifying the *logical structure* of the information, and a Web based platform consisting of a system that allows exploration of a public budget through interactive visualizations.

Finally, I examine public deployments of the system within the government of Uruguay, and the city of Bahía Blanca (Argentina).
Background
Prior Work

To make decisions, governments often need to gather and analyze information. However, public availability government information is a relatively recent phenomenon: the first public records legislation was introduced by Sweden in the late 18th century (Mustonen, 2006). The introduction and subsequent democratization of digital information processing techniques have naturally prompted governments, and the civil society, to extend the principles of access to public information to the “raw” or unprocessed data records generated by the state, instead of aggregate statistics and prepared reports.

In 2009 the United States Government also begun pushing open data when President Barack Obama introduced the Open Government Directive (Orszag, 2009), which mandated “executive departments and agencies to take specific actions to implement the principles of transparency, participation, and collaboration”. This mandate institutionalized this demand for digitized public information, and nurtured the emergence of several movements which can be collectively described as civic technology.

In this context of readily available machine-readable datasets generated by governments, and a community of eager journalists, civic hackers and companies, the last few years saw an explosion of platforms that attempt to make often complex information more accessible to non-specialist audiences. Fiscal (budget, expenditure and revenue) information is a natural candidate for this kind of tools, since management of public money has been historically associated with transparency and oversight of the activities of the government.

Although there are precedents of governments publishing contracting and expenditure information online as early as the year 2000\(^1\), those initial attempts were simple “data dumps” from the administrations' financial management information systems (FMIS),

---

\(^1\) The city of Bahia Blanca (Argentina) implemented the Active Control system in September, 2000: http://www.bahiablanca.gov.ar/digesto/Ordenanza1.html?ord=11162
which often failed at making the fiscal information accessible to the general public and public servants. One of the earliest and most referenced examples of a platform that made use of interactive data visualizations is *Where does my money go?* (Gray & Chambers, 2012), originally launched in 2009\(^2\) by the UK non-profit Open Knowledge Foundation. This project evolved into Open Spending, an openly accessible data visualization platform with features tailored to datasets pertaining to fiscal information.

![Image of Where does my money go? platform](image)

*Figure 2: Where does my money go? One of the first examples of data visualization applied to fiscal data*

Commercial vendors also offer tools for visualization of budget and spending data, prompted by the *Open Government* and *Open Data* movements originated in 2007 and

---

institutionalized by the aforementioned presidential memorandum. The Seattle-based company Socrata sells Open Budget, a platform adopted by several districts in the US. Another commercial offering is OpenGov, which also provides governments with a dashboard for the display of fiscal information.

![Image of Socrata Open Budget](image)

*Figure 3: Socrata Open Budget showing City of Cambridge's education expenditure*

Besides the tools originated from the non-profit and commercial sectors, some governments have chosen to develop these tools in-house. A common issue with government-built platforms is that they do little to make fiscal information more accessible. Frequently, the tools that governments make available to the public are trimmed-down versions of the business intelligence tools that are used internally. A comprehensive survey
of government fiscal information portals can be found in (Dener & Min, 2013).

News organizations also developed tools to assist in the interpretation of budgetary data. The New York Times' *Four Ways to Slice Obama's 2013 Budget Proposal*, highlights multiple perspectives of the multi-dimensional nature of budget datasets.

*Figure 4: New York Times visualization of the 2013 US Budget proposal*

My own past experiences with information tools for fiscal information also influenced the design and implementation of *SpendView*. *Gasto Público Bahiense (Bahía Blanca Public Spending)* (Aristaran, 2010) is a site that displays procurement information from the government of Bahía Blanca, my hometown. *Presupuesto Abierto* (Aristaran, 2015), a project made as part of the initial explorations for the research presented on this thesis, is a system for graphical display of fiscal information, developed in collaboration with the
government of the City of Bahía Blanca (Argentina).

Figure 5: PresupuestoAbierto.org displaying fiscal information for the cities of Boston (USA) and Bahía Blanca (Argentina)

These examples show that dashboard systems are a frequent choice of paradigm for visualizing fiscal information. We define dashboard as a set of graphical representations of different perspectives of the data and interface elements that, similarly to a car or the cockpit of an airplane, allow the user to obtain knowledge at a glance about the state of the system, described by the measurements contained in the data.

I believe that the dashboard paradigm is not always well suited to the interaction conventions of the World Wide Web. Namely, they don't adapt well to low resolution
displays (e.g., mobile devices), they don't provide hyperlinks to individual configurations of filters, don't take advantage of design elements of linear narrative, and are usually “invisible” to search engines.

The Relevance of Dimensional Modeling

The dimensional modeling technique (Kimball & Ross, 2002) allows users to model knowledge about a system using the concepts of orthogonal dimensions and measures. Dimensions are the structural attributes of the system of interest, which belong to a similar category in the user’s perception of data. Measures are the quantitative observations about the system. For example, for a business that sells products in multiple markets and is interested in tracking its performance over time, the dimensions of interest will be: product, market and time, while the only measure in the model will be the price at which the products are sold. Additionally, dimensions can be structured hierarchically: in the previous example, products can be grouped in categories, markets according to their geographical location, while the chronological dimension can be analyzed by year, quarter and month.

Dimensions and measures are then linked together in a data cube, which becomes the central entity of a data warehouse system. The technologies that enable the implementation of these systems are known as Online Analytical Processing (OLAP). In contrast to Online Transaction Processing (OLTP), which support operational tasks such as recording individual transactions in an organization, OLAP systems are designed and optimized for executing complex, aggregated queries.

As an example, each record in UN COMTRADE (United Nations, 2014)—a database of international trade between countries that collects data since 1962—contains information
about the monetary value of imports and exports between a country of origin and a country of destination, recorded by year and by product, which are classified according to a predefined hierarchy.

The *entities* (time, countries, products, import/export amount) of this database can be mapped directly to concepts in the *dimensional* modeling paradigm as follows:

**Dimensions**

- **Chronological**
  - Year

- **Origin Country (Geographical)**
  - Continent
  - Country

- **Destination Country (Geographical)**
  - Continent
  - Country

- **Products**
  - Classification level 1
  - Classification level 2
  - Classification level 3

**Measures**

- Imports amount
- Exports amount
Users of this database may never need to know that in 2015 Argentina exported $13.8 millions of US dollars’ worth of oranges to Greece (an individual fact in the database), but instead what was the trade balance between those two countries in the last 15 years. With the information arranged in the aforementioned cube in an OLAP system, the user can apply operations to the data structure to obtain the desired information:

- **Slice** (filter) the cube along the Origin Country dimension, choosing the value “Argentina”
- **Slice** (filter) the cube along the Destination Country dimensions, choosing the value “Greece”
- **Drill down** to the “Year” level in the Time dimension
- **Roll-up** the Imports and Exports measures, applying the summarization rule:

  \[ \text{Balance} = \text{Exports} - \text{Imports} \]

In terms of implementation, multidimensional models are often materialized in relational databases, but their table structures differ substantially to the conventional *third normal form (3NF)* (Codd, 1972) found in OLTP systems. 3NF databases are usually structured following the *Entity-Relationship Model* paradigm (Chen, 1976), that seeks to reduce redundancy in the representation of entities and the relationships among them, with the goal of optimizing the system for a high volume of individual transactions, as each operation will involve the minimum possible number of modifications. The structure of a database schema suitable for an OLAP system allows for redundancy, as it does not need to support a high volume of individual modifications. In contrast to 3NF structures, they’re known as ‘star schemas’, examples of which are shown in Figure 6 and Figure 7.
Multidimensional Modeling of Fiscal Information

Fiscal information, revenue and expenditure, is recorded under different classifications, that are independent to each other. Each line of a public budget is usually imputed to at least 4 different hierarchical classifications, namely administrative, economic, functional, and chronological (Jacobs et al., 2009). The administrative classification identifies the area of government that is responsible for administering the funds, while the economic one records the type of expenditure incurred, for example, salaries, goods and services, interests or capital spending. The functional classification codifies the purpose and objectives of the spending, such as education, public safety or community development. The latter has been standardized by the United Nations in the Classification of the Functions of Government (COFOG) manual, to facilitate comparisons between government administrations (Department of Economic and Social Affairs, 2000), while the economic one was codified in the Government Finance Statistics Manual by the International Monetary Fund (International Monetary Fund, 2014).

Some government administrations use additional classification criteria. Among them, we can identify the geographic classification, that identifies the region in which the revenue was collected or the expenditure was incurred, as well as the funding source, that classifies the origin of the perceived or expensed monetary resources.

Associated to each line of the fiscal dataset, which contains references to entities the classifications in use, there are also monetary amounts, which correspond to the phases of the budget execution process. Typically, these are: proposed (budget proposal by the executive or legislative branches of government), approved (budget as approved by the legislative branch), adjusted (subsequent modifications during the fiscal year) and executed (actually perceived revenues or incurred expenses). Other administrations may choose to use additional phases such as committed or paid.
These mutually independent classifications facilitate multiple perspectives of analysis. An economist doing long-term planning could be interested in the yearly amounts broken down by the first level of the functional classification, while an auditor might want to obtain a detailed report of spending by week of a particular month, for a certain department.

<table>
<thead>
<tr>
<th>Chronological Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
</tr>
<tr>
<td>2015</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Administrative Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
</tr>
<tr>
<td>Ministry of Agriculture</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Economic Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section</td>
</tr>
<tr>
<td>Ordinary Expenses</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Functional Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Function</td>
</tr>
<tr>
<td>Social Services</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Additional Classifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source of financing</td>
</tr>
<tr>
<td>National Treasury</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expenditure Phases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
</tr>
<tr>
<td>$176,455,000 ARS</td>
</tr>
</tbody>
</table>

*Table 1: Example classification and phases for a line of an expenditure budget*

Table 1 illustrates how each classification has features that differ from and are independent of the others. In well-designed budget structures, each transaction is attributed unambiguously to each of the defined classifications.

This normative framework for public finances, now adopted by most governments, translates almost directly to the principles of dimensional modeling. As an example, the budget structure to which the line depicted in Table 1 belongs, would be represented in
an OLAP system as follows, in terms of dimensions, hierarchies, levels and measures:

**Dimensions**

- Chronological
  - Year
  - Month
- Administrative
  - Title
  - Section
  - Chapter
- Economic
  - Section
  - Article
  - Paragraph
- Functional
  - Main Function
  - Function
  - Secondary Function
- Source of Financing
  - Source of Financing
- Geographic
Measures

- Proposed
- Adopted
- Committed
- Paid

A database structure suitable for storage of a typical budget dataset is depicted in Figure 6.

*Figure 6: A "star schema" for a budget*
There is an increasing number of governments publishing detailed budget information that preserve the multi-dimensionality needed for proper analysis. However, the lack of a data standard that includes information (metadata) about the classifications, taxonomies and expenditure phases in use in a particular administration, is a hurdle for the adoption of effective visualization and analysis tools for public finances. The *Fiscal Data Package* (Björgvinsson, Pollock, & Walsh, 2016), a recent standardization effort led by the Open Knowledge Foundation, strives to provide users with such a standard. This format also realizes the usefulness of the dimensional modeling concepts for representing, storing and transmitting government budget information.
Implementation
Design Rationale

This section outlines the design principles behind SpendView, provides an overview of implementation, and describes current known limitations of the system.

Design principles

The design of SpendView was informed by a number of design principles, that attempt to improve on the dashboard paradigm as device for visualizing multidimensional datasets, choosing public budgets as the focus of this work.

Narratives

Stories are perhaps the primary mean for transmission of knowledge among humans. However, their usage as a device for extracting knowledge of a multidimensional dataset is not frequent. By structuring information in the form of the journalistic inverted pyramid, translating particular slices of the data cube into questions and their answers, and combining automatically generated graphical and linguistic summaries, SpendView aims to ease the comprehension of complex information.

Hypermedia

Although most current data exploration and visualization systems are distributed as Web applications because it is technically convenient, they don't use the language of the Web. These dashboard-type systems originated in workstations, and they retain pre-Web user interface conventions. SpendView aims to be not just an application that runs in a Web browser, but a platform that embraces the idioms of the Web, encouraging linking, embedding, indexing, archival and sharing of its content.
Internationalization

*SpendView* is conceived globally. Having automatically generated text as one of its main interaction devices, its user interface has been designed from the start with internationalization in mind. At its present state, it supports English and Spanish; supporting additional languages only requires creating a substitution file that contain translated strings.

Openness and reusability

*SpendView* and *Mondrian-REST*, the two components that comprise the system, are licensed under the MIT license, and are freely available on the code-sharing platform GitHub.

---

**Mondrian-REST: a Web-friendly API for OLAP**

This section briefly describes the *Multidimensional eXpressions (MDX)* language, and introduces a novel querying interface OLAP databases that support MDX.

Projects that implement tools for visualization of multidimensional data often build *ad-hoc* interfaces for querying it. These application programming interfaces (APIs) contain knowledge about the particular dataset that they allow to query, frequently encoded in imperative programming languages, with the corresponding loss of generality that prevents reuse of those software components. OLAP databases offer a powerful declarative mechanism for specifying the structure of the data, and the query language they provide is expressive enough to encode every possible aggregation of the data.

**Multidimensional eXpressions (MDX)**

The MDX language was introduced by Microsoft in 1997, as part of the OLE DB for OLAP specification. It provides a specialized syntax for defining and querying
multidimensional objects and data, such as the data stored in OLAP cubes. While similar to SQL, it is not limited to bi-dimensional tables (rows and columns); MDX allows the user to query structures of up to 127 dimensions, which are referred as axes.

Taking the UN COMTRADE cube described in a previous section, a user would issue the following query get the yearly total amount of exports from Argentina to each of the African countries:

```sql
SELECT {[Measures].[Exports]} ON Axis(0),
        [Time].[Year].Members ON Axis(1),
        [Destination Country].[Africa].Children on Axis(2)
FROM [Trade Flow]
WHERE ([Origin Country].[South America].[Argentina])
```

This query will select the Exports value —which is part of the special Measures dimension—in the first axis, which by default is aggregated using the sum operator. The second access will be populated with the contents of the Time dimension, at the Year level. The third and last axis will contain the immediate descendants (Children) of the Africa member of the Destination Country dimension. The data that will take part on the aggregation will be those records (tuples) for which the Origin Country dimension is equal to Argentina, as specified by the WHERE clause. Table 2 shows an excerpt of the result of the query.

<table>
<thead>
<tr>
<th>Year</th>
<th>Country Code</th>
<th>Country Name</th>
<th>Exports</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>gab</td>
<td>Gabon</td>
<td>2326825</td>
</tr>
<tr>
<td>1995</td>
<td>gha</td>
<td>Ghana</td>
<td>2556622</td>
</tr>
<tr>
<td>1995</td>
<td>gin</td>
<td>Guinea</td>
<td>436794.32</td>
</tr>
</tbody>
</table>
A REST endpoint for OLAP databases

Despite the flexibility and expressive power of MDX and multidimensional databases, there hasn't been significant adoption of these technologies outside the enterprise. Analytical databases that support MDX integrate with other systems through protocols like the Microsoft Windows-only OLE DB for OLAP or XML for Analysis (Microsoft, 2001). However, the lack of a protocol that conforms to current conventions and patterns for Web-applications might have contributed to the low levels of adoption of multidimensional databases. This section introduces Mondrian-REST, an implementation of an URL-encoded query language that is able to express most common types of MDX queries.

The design of the proposed query language is centered around 3 operations:

- **Drilldown:** allows to move from one level of detail to the next. Starting from the root of the UN COMTRADE data cube, drilling down on the Origin Country dimension means getting information for every Continent. Multiple drilldowns can
be specified in a single query.

- **Cut**: allows the user to specify a filter, which effectively restricts the cube in which the aggregations are performed. *Cutting* the cube along the `[South America].[Argentina]` member in the Origin Country dimension, will only consider cells for which Argentina is the origin country in the trade data. Multiple cuts can be specified in a single query.

- **Measures**: allows to select the measures (scalar variables associated with a particular *fact* in the data cube). Multiple measures can be selected in a single query.

These three operations are sufficient to express the most common types of queries needed by sites like *SpendView*.

Additionally, Mondrian-REST provides access to the structure of the cube, allowing the user to obtain metadata such as the list of members of a particular dimension, or detailed information about a member.

The following examples illustrate how queries are constructed, by restricting the search space with the *drilldown*, *cut*, and *measures* parameters.

The single entry point for aggregation operations in Mondrian-REST is `/aggregate`, which accepts the described parameters. In order to aggregate (sum) the *Imports* measure over the entire cube (every country), a user would issue the query:

```
.aggregate?measures[]=Imports
```

To obtain the time series of the total imports of every country, a *drilldown* on the Time dimension must be performed:

```
.aggregate?drilldown[]=Year
```
To add an additional dimension, it is possible to specify an additional drilldown. When combined with the `cut` parameter, this query adds every African country to the dataset, resulting in a yearly time series of total imports by each African country:

```
/aggregate?drilldown[]=Year
  &drilldown[]=Destination Country
  &cut[]=Destination Country.Africa
  &measures[]=Import
```

Finally, this query will only consider imports of live horses by African countries, originated in South American countries.

```
/aggregate?drilldown[]=Year.Year
  &drilldown[]=Destination Country
  &cut[]=Origin Country.South America
  &cut[]=Destination Country.Africa
  &cut[]=Product.Live Animals
  &measures[]=Imports
```

The default representation of the query result is a JSON-encoded dictionary, that contains information about the members of each dimension along with the associated measures. Additionally, following the principles of Representational State Transfer (REST) (Fielding & Taylor, 2000), now adopted as the standard in modern Web application development, the user can also request that the response is encoded as comma-separated values or as an Excel spreadsheet. These tabular (two dimensional) representations of a multidimensional structure follow the “tidy data” recommendations (Wickham, 2014) that allow for easy processing and manipulation with standard data analysis tools.
Within Mondrian-REST, the mapping between logical entities (dimensions and measures) and the underlying "physical" representation of the information in the database is achieved through a configuration file in XML format (manually created by the developer), that encodes the principles of dimensional modeling. A fragment of such a file for the UN COMTRADE database schema (Figure 7) dataset is shown in Figure 8.

\[
\begin{array}{|c|}
\hline
<table>
<thead>
<tr>
<th>dim_time</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
</tr>
<tr>
<td>year</td>
</tr>
</tbody>
</table>
\hline
\end{array}
\]

\[
\begin{array}{|c|}
\hline
| dim\_country |
| id |
| id\_continent |
| continent |
| id\_country |
| country |
\hline
\end{array}
\]

\[
\begin{array}{|c|}
\hline
| facts\_trade |
| id |
| id\_time |
| id\_origin\_country |
| id\_destination\_country |
| id\_product |
| value\_imports |
| value\_exports |
\hline
\end{array}
\]

\[
\begin{array}{|c|}
\hline
| dim\_products |
| id |
| id\_category\_level\_1 |
| category\_level\_1 |
| id\_category\_level\_2 |
| category\_level\_2 |
| id\_category\_level\_3 |
| category\_level\_3 |
\hline
\end{array}
\]

Figure 7: Database "star schema" for UN COMTRADE
The preceding examples use the UN COMTRADE dataset to demonstrate the expressiveness of the multidimensional modeling paradigm and the proposed query language. These techniques were applied to fiscal information during the implementation of the systems presented in this thesis. Using the model for a budget dataset listed in page 30 and the budget line of Table 1 as examples, the following query returns a yearly time series of all the expenditure phases, broken down by the first level of the functional classification, for the Ministry of Agriculture:
Implementation Details

Mondrian-REST is an HTTP service that translates queries encoded in the previously described language into MDX, which is in turn interpreted by the Mondrian Relational OLAP (ROLAP) engine\(^3\). This software component translates MDX into SQL queries, which are executed by a relational database. Mondrian ROLAP supports most relational databases: the system built for this thesis used the lightweight relational database system SQLite\(^4\) during the development process, and the column-oriented database MonetDB\(^5\) for live deployment with no change in the underlying code. The system has also been successfully deployed using the PostgreSQL\(^6\) and MySQL\(^7\) database systems.

Mondrian-REST is implemented in JRuby\(^8\), a version of the Ruby language that runs in the Java virtual machine.

\(^3\) [http://community.pentaho.com/projects/mondrian/](http://community.pentaho.com/projects/mondrian/)
\(^4\) [https://sqlite.org/](https://sqlite.org/)
\(^5\) [https://monetdb.org/](https://monetdb.org/)
\(^6\) [http://postgresql.org/](http://postgresql.org/)
\(^7\) [https://mysql.com/](https://mysql.com/)
\(^8\) [http://jruby.org/](http://jruby.org/)
User experience and user interface

*SpendView* structures the overview of a fiscal information dataset linearly, inspired by the well-known design concept of a *profile page* and the *inverted pyramid* writing technique of journalistic articles. This writing style is characterized by “the most important information summarized in the so-called ‘lead sentence’ that, according to standard practice, has to answer four or five ‘w-questions’ (who? when? where? what? and perhaps why?). After the lead sentence comes the rest of the story [...]” (Pöttker, 2003).

In addition to interactive visualizations, *SpendView* relies in automatically generated textual narrative that describes and highlights basic statistics of the datasets.

This section provides an overview of the UI and describes in detail how the user interacts with the system to explore a budget.

Overview of a Budget

The initial view of a district’s budget is shown in Figure 11. Following a modular structure, the page contains two types of elements.

The header (Figure 9) contains the name of the district and its logotype or emblem. It also contains a search input field, that allows the user to search for budget entries. It also contains a brief textual summary of some aspects of the budget, that include the total amount for the latest available year, its *compound annual growth rate* and the main sources of expenditure for all the available budget classifications. Following the header, the page contains one *aggregation view* module (Figure 10) for each classification of the budget. By default, it displays the first two levels of a dimension in a *treemap*, a visualization technique suitable for hierarchical data structures (Johnson & Shneiderman, 1991).
In 2016, City of Cambridge's budget was of 546 million. It has increased at an annualized rate of 3.5% since 2011.

According to the administrative classification, the main areas of spending were Education (164 million, 30%), Debt Service (54.7 million, 10%) and Police (50.6 million, 9%).

According to the economic classification, the main areas of spending were Salaries and Wages (365 million, 67%), Other Ordinary Maintenance (117 million, 21%) and Extraordinary Expenditures (54.7 million, 11%).

According to the functional classification, the main areas of spending were Education (164 million, 30%).

According to the funding source classification, the main areas of spending were General Fund (151.9 million, 29%), Water Fund (14 million, 3%) and Parking Fund (11.5 million, 2%).

Figure 9: Detail view of the header, showing the generated textual description of the data

Figure 10: An "aggregation view" showing the first 2 levels of the administrative classification
Figure 11: SpendView showing the initial overview of the budget for the City of Cambridge
Drilling down on a budget classification

The *drill down* operation allows the user to move from one level of detail to the next. *SpendView* maintains the design of the main page for displaying information about a *member* of a particular budget classification. Figure 12 shows the profile for the “personal permanente” (permanent staff) classification—a second-level classification in the administrative dimension, within the “gastos en personal” (salaries and wages) item—in a municipal budget. In addition to the elements of the main budget view described in page 45, the heading contains a *trail* that indicates the depth of this member within the overall budget, along with the names of the dimension and hierarchy level in which the member is located.

![Figure 12: Partial view of a member of a budget](image-url)
Exploring the different perspectives of the budget

The container for most interactive elements in SpendView is the “aggregation view”, shown in Figure 10. Both the budget view (Figure 11) and drilldown pages (Figure 12) include one of these modules for each available dimension of a budget.

Budget dimensions as questions

As mentioned in page 18, budgets are classified according to multiple, independent dimensions that offer different perspectives of analysis. The usual nomenclature to these classifications can be translated to more accessible language, to facilitate comprehension of the budgetary information (Table 3).

<table>
<thead>
<tr>
<th>Question</th>
<th>Classification Name</th>
<th>Example members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who spends?</td>
<td>Administrative</td>
<td>Police department, Ministry of Education</td>
</tr>
<tr>
<td>What is spent on?</td>
<td>Expense Type</td>
<td>Salaries and Wages, Goods, Services</td>
</tr>
<tr>
<td>What is the funding source?</td>
<td>Financing source</td>
<td>General Fund, Revenue Tax, External Financing</td>
</tr>
<tr>
<td>What is the purpose?</td>
<td>Functional</td>
<td>Community Development</td>
</tr>
<tr>
<td>Where it is spent?</td>
<td>Geographic</td>
<td>Jamaica Plains, Buenos Aires Province</td>
</tr>
</tbody>
</table>

Table 3).
<table>
<thead>
<tr>
<th><strong>What is spent on?</strong></th>
<th><strong>Expense Type</strong></th>
<th><strong>Salaries and Wages, Goods, Services</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What is the funding source?</strong></td>
<td><strong>Financing source</strong></td>
<td><strong>General Fund, Revenue Tax, External Financing</strong></td>
</tr>
<tr>
<td><strong>What is the purpose?</strong></td>
<td><strong>Functional</strong></td>
<td><strong>Community Development</strong></td>
</tr>
<tr>
<td><strong>Where it is spent?</strong></td>
<td><strong>Geographic</strong></td>
<td><strong>Jamaica Plains, Buenos Aires Province</strong></td>
</tr>
</tbody>
</table>

*Table 3: Budget classifications as translated to questions*
Data verbalization as a complement of visualizations

*SpendView* makes use of automatically generated linguistic summaries of the data that is used to construct budget and dimension member profiles. Each aggregation view contains a portion of text that describes the members of the displayed dimension that incurred in the larger portions of expenditure, according to the displayed classification.

For example, given the budget of the City of Cambridge for the fiscal year 2016, the system would generate the following summary for the *functional* ("What is the purpose?") classification:

*In 2016, Education (164 million USD, 30%) and Public Safety (120 million USD, 22%) were the two main areas of spending in the functional classification.*

On a dimension member profile, the text is expanded to indicate that the summary is limited to that particular cell. The *Education* profile for the budget of Cambridge contains this text in its *economic* ("What is it spent on?"") classification section:

*In 2016, Salaries and Wages (137 million USD, 83%) and Other Ordinary Maintenance (25.4 million USD, 16%) were the two main areas of spending in the economic classification within Education.*

Additionally, the names of the dimension members that are included in the linguistic summary are hyperlinked to their respective profiles.

These linguistic summaries are crucial for capturing traffic originated in search engines. Charts present data in graphical formats, but they are effectively “invisible” to search engines. Automatically generated text, on the other hand, is quickly detected and indexed by search engines (Figure 18).

Multiple graphical representations

The default graphical representation of the data contained in an aggregation view module
is the treemap. According to the dimension being displayed, the treemap will represent up to two levels in the hierarchy; the first level is encoded by a color in a categorical scale, and the second is encoded by the position of the rectangle that represents a member, as calculated by the Squarified algorithm (Bruls, Huizing, & van Wijk, 2000). Treemaps are effective for representing the relative share or importance of items in a total amount. In SpendView, the treemap items (rectangles) are budget categories in a classification, and their size is relative to their contribution to the overall spending in a particular year. The user can select a year of interest with a “slider” control situated below the treemap. However, this is not effective for visualizing temporal trends, a frequent use case in budget analysis. To accommodate for this, SpendView offers line and stacked bar charts, that represent the temporal dimension on their horizontal axis (Figure 13).

Exploring the data
Every chart type in SpendView is interactive, allowing the user to obtain more information about the displayed budget categories. When the mouse is over either a treemap rectangle, a line in line chart, or a bar in a barchart, it presents an indicative tooltip that contains detailed information about the budget classification. Figure 14 shows a tooltip displaying the amounts of the execution phases of a budget, in the selected year. Chart legends also react to the mouse pointer, adding to the tooltip’s content, controls for hiding or isolating that particular series from the chart.
In addition to the "mouse over" interaction gesture, charts and legends also respond to clicks. This interaction is used to obtain more information about a budget item. When
the user clicks either on a chart element or on an item of a chart's legend, SpendView displays a modal window constrained to the aggregation view module that contains the chart. This modal window introduces an additional visualization method, the grouped bar chart, that facilitates comparisons of multiple expenditure phases (e.g., approved, adjusted, executed). It also utilizes a tooltip, consistently with the main aggregation view, to provide further details.

![Figure 16: Modal window showing spending time series for a budget item](image)

Sharing
Ease of sharing is one of SpendView's design principles. In addition to every budget and item page having an unique URL ("permalink"), each aggregation view module offers multiple ways to share and download the displayed information in several formats. An "options" button displays a modal window (Figure 17), similar to the one described in the preceding section, that contains the following options:
• **Share:** buttons to share the selected view on social media platforms.

• **Download:** buttons to download the data corresponding to the current chart in comma-separated (CSV) and Microsoft Excel (XLS) formats. The user can also download a rendering of the chart in Portable Network Graphics (PNG) and Scalable Vector Graphics (SVG).

• **Data:** Tabular representation of the data.

• **API:** Displays the address of the application programming interface (API) for consuming the information through the HTTP protocol.

• **Embed:** HTML code suitable for embedding the selected visualization into a Web page.

![Figure 17: The Share modal window displaying the "Data" section](image)

**Implementation of the user interface**

*SpendView* is currently running at spendview.media.mit.edu as a public Website with access to the open Web. The following sections discuss how the interface was
implemented.

**Isomorphic Web applications**

Web applications, being client-server systems, are comprised of two parts: the server application, that serves content (markup, media, executable code) to a Web browser through the network to a client; and the client application that runs on the user’s device. Before the advent of rich interfaces, these applications were simple in scope. The HTTP served HTML files and media content to a Web browser, that rendered them and only provided basic interactions such as navigating through hyperlinks, or submitting the contents of a form to a server.

In 1995, Netscape introduced the interpreted language *Javascript* to its Navigator product, which enabled the creation of richer interactions in the Web browser by manipulating the contents of the HTML that was downloaded from the server, in response to the actions of the user. These new possibilities lead to an increase in the complexity of Web applications, which now consist of two separate codebases —server and client—, usually implemented in different programming languages. Furthermore, less-capable HTTP clients —such as the crawlers employed by search engines to gather and index the contents of Web pages— are unable to execute Javascript code. In consequence, information that is displayed on a Web browser screen by manipulating HTML with Javascript, is in effect invisible to search engines.

In response to this, and with the recent feasibility of Javascript as a server side language (Tilkov & Vinoski, 2010), the Web development landscape saw the emergence of a new development practice, called *Isomorphic* or *Universal* Web applications (Robbins, 2011). This technique allows the developer to employ a single, coherent technology stack for the server and client applications. *SpendView* is developed with such a paradigm: the entire
codebase for the Web application is developed with Javascript, using the React application framework, released by Facebook in 2014.

Visibility to non-rich internet clients
One of the main benefits of rendering HTML content on the server, as mentioned above, is the increased “visibility” to search engines (Figure 18) and to social media platforms. This is particularly relevant in a context where social sharing and organic search results amount to a combined 80% percent share of the sources of traffic acquisition for websites. Social media platforms have created mechanisms to increase visibility of shared content, which require the inclusion of metadata tags that are interpreted by the platform’s crawlers. SpendView uses these devices to provide a textual summary of the information contained in the link that was shared.

Figure 18: Organic search results in Google for the query "spendview argentina"

9 https://facebook.github.io/react/
Playing with @CityOfBoston’s open data:

SpendView: City of Boston
In 2016, City of Boston’s budget was of 2,857,412,184. It has increased at an annualized rate of 4.61% since 2013.
spendview.media.mit.edu

SpendView: República Argentina
In 2016, República Argentina’s budget was of 1,564,422,349,338. It has increased at an annualized rate of 20.89% since 2008.

Figure 19: Sharing on Twitter

Figure 20: Sharing on Facebook
Impact
Bahía Blanca: from confrontation to collaboration

Bahía Blanca, a city with a population of 300,000 in the south of the Buenos Aires province in Argentina, was among the first in the world to use the Internet to publish detailed fiscal information: in the year 2000, its government implemented a Web site that let citizens consult the city’s checkbook. The system, although primitive for today’s standards, allowed users to obtain information on most aspects of the city’s revenues, expenditure and procurement. The successive administrations trimmed down the system considerably; in mid-2010 it only remained as a simple list of purchase orders emitted by the government.

In July 2010, motivated by some corruption scandals that were reported by the local media, I built GastoPublicoBahiense.org\(^{10}\) (GPB), a system that obtained those procurement records through screen scraping techniques and produced a Web site that republished that information using interactive visualizations (Figure 21). This initiative gave relevance to a dataset that had been available for more than 10 years, and had immediate impact on the city’s political and media landscape. It is also frequently cited as a pioneering example of civic hacking in Latin America (Jefatura de Gabinete de Ministros, 2014; Morozov, 2014).

The then-current administration took measures against the project, such as introducing a CAPTCHA—a challenge-response mechanism destined to block screen-scrappers—in section of the city’s Website where the procurement information was published. This incident brought more notoriety\(^{11}\) to the GPB project. It also prompted the next administration to create an office within government dedicated to transparency and

\(^{10}\) http://gastopublicobahiense.org
technological innovation.

In this new political context, the city’s government removed the previously established obstacles for accessing public information, and we established a collaborative relationship. In June 2015, in the context of this research, I started collaborating with Bahia Blanca’s government to develop a budget information portal for the city. The city has an integrated financial management information system (IFMIS) (United States Agency for International Development, 2008) in operation since 2008, but didn’t have a publicly-facing system for providing budgetary information to its citizens.

The outcome of this collaboration was twofold: I reverse engineered the IFMIS’s database and wrote a set of software components that allow the extraction of detailed budgetary

Figure 21: The homepage of GastoPublicoBahienne.org in July 2010
and procurement reports in machine-readable formats, suitable for use in data analysis tools. I also developed *Presupuesto Abierto* (Figure 5), an initial prototype for a budget information portal, that was publicly deployed in November 2015 and is running at http://bahia.presupuestoabierto.org.

The usage data and feedback collected from the public deployment of *Presupuesto Abierto* informed the development of *SpendView*, and drove the decision to explore alternatives to the dashboard paradigm. Additionally, the software that was developed to extract the financial records from the city’s IFMIS —a software system provided by the provincial government which runs in more than 140 districts across Argentina— was made available by Bahia Blanca’s government to other municipalities, which are now using it to overcome this system’s limited reporting capabilities, by processing this information with more powerful tools.

In April 2016, Bahia Blanca’s fiscal records from 2008 to 2016 were made available in *SpendView*. In contrast with the other budget datasets that are published in the platform, Bahia Blanca’s are kept updated thanks to the integration with the city’s IFMIS that was developed for this research.

Given the standardization and architectural homogeneity of IFMIS (United States Agency for International Development, 2008), it is feasible to develop integrations of *SpendView* with other systems.
Working with Uruguay's budget office

In March 2016, I spent a week working at the Uruguayan presidency’s Office for Budget and Planning (OPP)\(^\text{12}\), through a collaboration with the Latin American Initiative for Open Data (ILDA)\(^\text{13}\). The Uruguayan national government has one of the most active Open Data programs in Latin America\(^\text{14}\), and has maintained a citizen-oriented budget information portal since 2011 (Figure 23).

---

\(^{12}\) http://www.opp.gub.uy/

\(^{13}\) http://idatosabiertos.org/

\(^{14}\) Currently occupying the 19\(^{\text{th}}\) place in the world in the Open Data Barometer, out of 92 countries.
OPP is currently undertaking the renovation of this online resource, and has expressed interest in adopting *SpendView*, as it presents some advantages over their current portal. Namely, it allows for easier visualization of temporal trends, and enables queries *across* all the classification criteria (dimensions) of the budget. OPP also valued the *linguistic summaries* that are presented alongside charts.

**Workshop**
I conducted a workshop titled “Exploring technology and open budget data”, in collaboration with members of ILDA and OPP. It was attended by members of several ministries (Labor, Interior, Housing), consultants, journalists, government planners and
researchers from local universities. We structured the activity as a focused conversation: the participants were separated in groups, and were tasked with collaboratively producing answers to a series of questions posed by the organizers of the workshop, to be later shared and discussed. The following are some of the questions that we posed to the participants, and the answers that were produced:

*What are the challenges that you face when working with budgetary information?*

For the most part, participants pointed out deficiencies in the existing information resources: lack of performance-oriented indicators that can be associated with expenditure data, and the need for transaction-level information (checkbook entries) that would allow analysts to construct their own budget classifications. They also mentioned the lack of detail in the formulation of the country’s budget.

*Are the fiscal information resources accessible and easy to consult?*

Participants from the government, noted that it is often difficult to obtain budget information from government systems. Notably, some of them were unaware of the official budget information portal, and mentioned that it was not possible to reach it from a search engine query.

*How do you use budget information? For which purpose?*

Among others, participants from the governments and news organizations mentioned the use of budget data for creating reports to non-specialized audiences, performing comparative analysis of spending across government programs. Researchers and consultants referred to their use of budget records to study the effectiveness and cost of public policies.

**User evaluation of SpendView**

I conducted an informal testing session of SpendView with the participants of the
workshop. The tool was well-received, and I obtained valuable feedback and observations:

- Users valued that the system “can be read and navigated like a Web page”, in contrast to the business intelligence tools that they are used to.

- Although the linguistic summaries were perceived as useful, some users noticed the uniformity in their structure. Verbalizing other statistical properties of the information—besides top spenders—is intended to be developed in future versions of SpendView.

- Some users were confused by the visualization of two levels of a hierarchy on a single treemap. Future versions of the platform will display a single level by default, and provide a “depth” control, that allows users to control how many levels are displayed by the treemap.

**Public Deployment**

*SpendView* was deployed to the open Web on April 22, 2016, with 7 initial budget datasets (Table 4) loaded into the platform, totaling 1.6 million individual records.

<table>
<thead>
<tr>
<th>District</th>
<th>Type</th>
<th>Period</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uruguay</td>
<td>Country</td>
<td>2011-2015</td>
<td>Uruguay Open Data Portal(^{15})</td>
</tr>
<tr>
<td>Bahía Blanca</td>
<td>City</td>
<td>2008-2016</td>
<td>Bahía Blanca IFMIS</td>
</tr>
</tbody>
</table>

\(^{15}\) https://catalogodatos.gub.uy/dataset/credito-presupuestal-asignado-y-ejecutado-con-aperturas
<table>
<thead>
<tr>
<th>Location</th>
<th>Type</th>
<th>Period</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambridge, MA</td>
<td>City</td>
<td>2011-2016</td>
<td>City of Cambridge Open Data Portal(^\text{16})</td>
</tr>
<tr>
<td>Buenos Aires</td>
<td>City</td>
<td>2013-2015</td>
<td>Buenos Aires Open Data Portal(^\text{17})</td>
</tr>
<tr>
<td>Argentina</td>
<td>Country</td>
<td>2008-2016</td>
<td>Argentina’s &quot;Citizen’s Portal&quot;(^\text{18})</td>
</tr>
<tr>
<td>Boston, MA</td>
<td>City</td>
<td>2013-136</td>
<td>City of Boston Open Data Portal(^\text{19})</td>
</tr>
<tr>
<td>Poland</td>
<td>Country</td>
<td>2010-2014</td>
<td>World Bank’s BOOST project(^\text{20})</td>
</tr>
</tbody>
</table>

\(^{16}\) [http://budget.data.cambridgema.gov/](http://budget.data.cambridgema.gov/)

\(^{17}\) [http://data.buenosaires.gob.ar/dataset/presupuesto-ejecutado](http://data.buenosaires.gob.ar/dataset/presupuesto-ejecutado)


\(^{19}\) [http://budget.data.cityofboston.gov/](http://budget.data.cityofboston.gov/)


---

Table 4: Budgets loaded into SpendView as of April 2016

This initial set of budgets, that originate from heterogeneous sources, served to test one of the main design principles of the system, namely, that it should be possible to map the logical organization of a budget (hierarchical classifications and expense phases) with their physical representation (i.e., the data as stored in the database).

Table 5 shows the columns present in Boston’s budget dataset, as available for download in the city’s open data portal, and their mapping to budget classifications.

<table>
<thead>
<tr>
<th>Column</th>
<th>Classification</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiscal Year</td>
<td>Chronological</td>
<td>1</td>
</tr>
<tr>
<td>Cabinet</td>
<td>Administrative</td>
<td>1</td>
</tr>
</tbody>
</table>

16 http://budget.data.cambridgema.gov/

17 http://data.buenosaires.gob.ar/dataset/presupuesto-ejecutado

18 http://sitiodelciudadano.mecon.gov.ar/

19 http://budget.data.cityofboston.gov/

20 http://wbi.worldbank.org/boost/country/poland
Table 5: Columns and their mapping to budget dimensions

<table>
<thead>
<tr>
<th></th>
<th>Administrative</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>Administrative</td>
<td>3</td>
</tr>
<tr>
<td>Expense Type</td>
<td>Economic</td>
<td>1</td>
</tr>
<tr>
<td>Expense Category</td>
<td>Economic</td>
<td>2</td>
</tr>
</tbody>
</table>

This mapping is implemented with an XML file, similar to that shown in Figure 8, that specifies the columns that correspond to each classification of a budget, and the expenditure phases. At present, the maintainer of an instance of SpendView needs to manually edit this XML mapping. Implementing a graphical interface for this task is planned for a future version of the platform. However, the structural similarity of budgetary information, the flexibility of Mondrian-REST, and the dimensional modeling technique, proved to be sufficient for the purpose of implementing a general-purpose API and visualization engine.

Usage data
Since its public deployment in April 22, 2006, SpendView has had 998 unique users, 1,288 sessions and 2,244 page views, according to data collected with the Google Analytics platform. Social networks and organic search were the sources that originated the most traffic, accounting for 41.03% and 41.62% respectively. Although the traffic volume is still insufficient to make a comprehensive analysis, we recorded interesting queries by users that arrived to SpendView through a search engine:

- “ministerio de educación argentina presupuesto” (“ministry of educacion argentina budget”), which led the user directly to the corresponding profile page in SpendView.
In addition to tracking individual page views, user interface actions such as clicking on a chart element, changing the chart type, or opening a modal, were also recorded in Google Analytics. While the sample size is small, this information highlighted potential usability issues. For instance, a considerable portion of the sessions ended after the user clicked on a chart’s element that causes a modal window to be opened (Figure 16). This window is meant to provide detailed information about a budget item, and to provide a link to the profile of that item. Assuming that the users that clicked on the chart element corresponding to a budget item intended to get more information about it, the observation that they are likely to end their session at that point, suggests a deficiency in the design of that interaction.
Conclusion
Future Work

Based on the preliminary analysis of the usage data, received feedback, and public deployments, I identify possible opportunities for future improvements: a tool that eases the process of importing a new budget into SpendView could promote wider adoption. Specifically, the Fiscal Data Package standardization effort (Björgvinsson, Pollock, & Walsh, 2016), led by the Open Knowledge Foundation, provides a format that contains the data and metadata needed to construct a SpendView instance. Additionally, the reverse engineering of the IFMIS system running in Bahía Blanca that is mentioned in page 60, presents the opportunity to deploy SpendView to more than 140 districts in Argentina. While the initial results and observations are encouraging, more usability studies should be conducted to improve the user experience.

Concluding Remarks

SpendView is a platform for visualizing public budget information, that makes use of linguistic summaries and interactive visualizations, that emphasizes linear narratives and sharing of content. Although the design and implementation principles outlined in this thesis were applied to fiscal information datasets, they could also be applied to other domains, taking advantage of the flexibility of the dimensional modeling paradigm.
List of Figures

Figure 1: A traditional dashboard for government finances ........................................ 16
Figure 2: Where does my money go? One of the first examples of data visualization applied to fiscal data ........................................................................................................ 23
Figure 3: Socrata Open Budget showing City of Cambridge's education expenditure ...... 24
Figure 4: New York Times visualization of the 2013 US Budget proposal .................. 25
Figure 5: PresupuestoAbierto.org displaying fiscal information for the cities of Boston (USA) and Bahía Blanca (Argentina) ........................................................................ 26
Figure 6: A "star schema" for a budget........................................................................ 33
Figure 7: Database “star schema” for UN COMTRADE ........................................... 42
Figure 8: Cube definition for the UN COMTRADE database.................................... 43
Figure 9: Detail view of the header, showing the generated textual description of the data.................................................................................................................................. 46
Figure 10: An "aggregation view" showing the first 2 levels of the administrative classification......................................................................................................................... 46
Figure 11: SpendView showing the initial overview of the budget for the City of Cambridge ........................................................................................................................................ 47
Figure 12: Partial view of a member of a budget ........................................................ 48
Figure 13: Treemap, line and stacked bar chart applied to the same dataset. ............ 53
Figure 14: Tooltip over a chart.................................................................................. 53
Figure 15: Tooltip over a chart legend ........................................................................ 53
Figure 16: Modal window showing spending time series for a budget item ............... 54
Figure 17: The Share modal window displaying the "Data" section ............................. 55
Figure 18: Organic search results in Google for the query "spendview argentina" ...... 57
Figure 19: Sharing on Twitter .................................................................................. 58
Figure 20: Sharing on Facebook ............................................................................... 58
Figure 21: The homepage of GastoPublicoBahiense.org in July 2010

Figure 22: SpendView showing the profile for the 'Salaries and Wages' budget item (Bahía Blanca)

Figure 23: Official budget information portal (Presidency of Uruguay)
Index of Tables

Table 1: Example classification and phases for a line of an expenditure budget .......... 31
Table 2: Result from MDX query (excerpt) ............................................................... 39
Table 3: Budget classifications as translated to questions ........................................ 50
Table 4: Budgets loaded into SpendView as of April 2016 ...................................... 67
Table 5: Columns and their mapping to budget dimensions ........................................ 68
http://doi.org/10.1145/320434.320440
Datawheel, Macro Connections / MIT Media Lab, & Deloitte. (2016). Data USA.


