Publishing Open Election Data

Leigh Dodds, Open Data Institute, 5th August 2014

This report presents the results of a short research project exploring current practices relating to the publication, collection and sharing of election data. Included are some recommendations regarding the publication of election data as open data in a standard format.

The report is accompanied by a technical specification for both a simple tabular format and an RDF vocabulary that supports the publication of election results as open data. These formats use a common conceptual model which has been designed to be both simple and flexible enough support the variety of electoral systems used internationally.

The report is organised into the following sections:

- a review of the types of data that are collected and exchanged during elections
- a discussion of the benefits of publishing election results as open data
- a review of some existing projects and technologies that support publishing of election data
- a proposal for a standard model for election results

Types of Election Data

An election involves a number of inter-dependent processes all of which involve the creation and exchange of a variety of types of data. These can be broken down into three broad categories depending on when the data is collected and used:

- **Pre-Election Data** -- data that is collected before an election is carried out. This includes:
    - registration of voters
    - registration of political parties
    - registration of candidates
    - location and opening times of polling stations
    - definition of electoral regions
    - election metadata, e.g. timing
- **Election Data** -- the recording and counting of voter preferences during the electoral process
- **Post-Election Data** -- the reporting, analysis, and auditing of votes and the communication of the results of the election

As an increasing number of regions and countries adopt electronic voting systems, more of this data is “born-digital” and is therefore available in a machine-readable format. In fact much of the Pre-Election Data is required to help support these voting systems, e.g. to authenticate

votes, generate ballot interfaces, and to collect data for reporting. However many electoral processes still use paper based systems.

Election data is interlinked with "reference data" which adds important context. This reference data includes details such as:

- Geographical metadata used to identify where voting has taken place, e.g. for regional analysis or to scope the results of elections
- Descriptions of the political parties with which candidates are affiliated.

Analysis of election results often makes use of additional metadata derived from the above, e.g. regional demographics, previous voting records for a region and party, etc.

Open Election Data

Open data is associated with a number of broad benefits, including economic development and increased transparency and accountability for public services. It is in the latter area where open election data may have the most benefits.However not all election data should necessarily be published as open data.

Referring to the process model outlined above, openness is best adopted in the publication of pre- and post-election data. Greater access to data relating to registration of voters and candidates, as well as to election results will help:

- Support the democratic process by ensuring that voters are well-informed and are able to engage with the system, e.g. by finding their polling station and understanding the choices they make
- Bring transparency to electoral processes, something of importance in many developing regions that are adopting democratic processes
- Support analysis and reporting of election results to provide insight into social processes and impacts
- Drive improvements to the electoral system, e.g. by measuring use of polling stations

Several of these potential benefits were highlighted by a joint project between the ODI and Deloitte[1]. A recent Open Knowledge Foundation blog post illustrates how access to open data can also help identify errors in election results[2].

Some types of election data should not be published as open data. Voter privacy is a fundamental principle of all democratic processes and detailed voting data itself should clearly

[1] http://theodi.org/blog/a-step-forward-for-democratic-engagement
[2] http://dk.okfn.org/2014/05/28/open-electoral-data-reveal-errors-in-danish-electoral-results/

remain private. However, even around the recording and counting of votes there are certain types of anonymised data which may still be usefully published.

For example “footfall” data relating to physical polling stations might usefully inform decisions around the placing and staffing of stations. Data on use of different voting options (physical, postal, electronic), particularly by different demographics, might help inform decisions around the design and adoption of electronic voting systems.

In order to achieve these potential benefits it is important to ensure that election data is:

- available from an authoritative, primary source
- available on a timely basis, e.g. published immediately after results are announced
- released under an open licence
- published in standard formats to facilitate reuse
- published according to a standard model that will support both aggregation of data and customization to allow for regional differences in election processes

Current Approaches to Publishing Election Data

There are a number of ways in which election data is published to the web, in many cases the data is published directly by the electoral commission for a country. Yet for many citizens the main source of election results will be news and media coverage rather than primary sources. There are also several academic research projects that aim to aggregate election data on an international basis.

This section summarises some of the initiatives and technologies which are relevant to this report. These have been organised into several broad categories:

- **Primary Sources** -- a review of strategies taken by public sector bodies to publish electoral data
- **Secondary Sources** -- election data aggregations, compiled by academics or crowd-sourcing efforts
- **Technologies** -- standards and technical formats for publishing election data

This section concludes by collating a common set of issues that can be addressed by increased standardisation and open licensing of election data.

Primary Sources

Many public sector bodies publish electoral data, this typically includes:

- Dates of forthcoming elections
- Results of past elections, often with statistical analysis of the results

For example the UK Electoral Commission publishes an election result archive[3] in addition to summary information about those elections[4]. It also maintain a public register of political parties[5]. Election results are published as Excel spreadsheets containing result data aggregated at various levels, e.g. by county and region.

The Election Commission of India[6] publishes a similarly broad range of election data, including details of political parties, individual candidates and a comprehensive statistical analysis of the results of each election. However, while comprehensive, the data is published as PDF documents which greatly limit re-use.

Interestingly, despite the use of electronic voting systems, the Estonian National Electoral Committee[7] publishes very little data about elections. Results are provided as a simple HTML table without any machine-readable data.

The Maryland State Board of Elections[8] provides an example of regional reporting in the US. Again, a broad range of data is available about forthcoming and past elections including statistical summaries available as PDF and some raw data published as Excel spreadsheets.

Even acknowledging that there are regional and constitutional differences there is a wide variety of different ways in which election results are reported by primary sources. The Open Data Index provides a good comparison chart[9] highlighting issues around the availability of election result data internationally.

Of the 70 countries listed in the Index half have not published machine-readable data about election results. Those that do publish machine-readable data do so in a variety of formats. Another significant issue is that the data, even when available in a machine-readable form, is not openly licensed.

Aggregations

There are a number of projects that attempt to aggregate election data from around the world in order to facilitate comparative analysis. These range from personal projects, such as Election

[3] http://www.electoralcommission.org.uk/our-work/our-research/electoral-data
[4] http://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums
[5] http://www.electoralcommission.org.uk/find-information-by-subject/political-parties-campaigning-and-donations/political-party-registration
[6] http://eci.nic.in/eci_main1/ElectionStatistics.aspx
[7] http://www.vvk.ee/general-info/
[8] http://www.elections.state.md.us/
[9] https://index.okfn.org/country/dataset/elections

Resources on the Internet[10] through to sites such as the Electoral Knowledge Network[11] which provides high-level comparative data on the design and adoption of different types of electoral systems.

Many of these aggregations are created to support academic research. Varying degrees of normalisation have been applied to the collected data to organise it into common formats. Examples include:

- Election Passport[12]
- Constituency Level Elections Archive[13]
- Global Elections Database[14]
- Election Reports Archive[15]

While these services typically offer a more standardized way to access election data, they have their open problems:

- Coverage can be limited, e.g. to specific countries, types of election, or specific periods in history, often reflecting specific research goals
- Data is sourced from a variety of places, including other secondary sources, with varying degrees of rigour in fact-checking, raising questions about reliability
- Data is often limited to overall election results with limited regional breakdown within individual countries
- Data is not necessarily tied to standard identifiers, e.g. country codes, regions, political parties, etc.
- Data licensing is unclear

Many of these issues stem from the fact that the source data itself is often hard to acquire or is unclearly licensed, resulting in manual effort to collect and curate the data.

Technologies

There are several projects whose goals are to help improve the dissemination of election data through the creation of standard data formats and publishing tools.

The Open Election Data[16] project was aimed at supporting local government in the UK in publishing the results of local elections. The project defined a simple RDF vocabulary for election results which could then be used to markup data on council websites using RDFa. This

[10] http://www.electionresources.org/
[11] http://aceproject.org/
[12] http://www.electionpassport.com/
[13] http://www.electiondataarchive.org/
[14] http://www.globalelectionsdatabase.com/
[15] http://cdp.binghamton.edu/era/index.html
[16] http://openelectiondata.org/

made it relatively easy for councils to adopt Linked Data and for aggregators to easily harvest and aggregate the data.

The Voting Information Project[17] is a US focused project whose goal is to provide voters with improved access to electoral information. The project has defined XML and CSV data formats for collecting a wide variety of different types of electoral data, including pre-election details such as polling station locations, through to reporting of election results.

The Popolo Project[18] aims to use community consensus to build a set of open government data specifications which can be used internationally. By defining simple, easy to use standards the project aims to make it easier for civic developers to re-use data published by government sources. The standards have been designed to harmonize with various existing data vocabularies and support publication in a variety of data formats including JSON and RDF. The specifications include drafts for describing motions, voting events and votes cast by legislative bodies.

Election Markup Language is a more formal attempt to create an electoral data standard. Developed by the OASIS Electoral and Voter Services Technical Committee[19] the specification defines a standard generic model for election processes and XML message formats that support data exchange at all stages. The formats have been designed to be generic so that they can be customized and used in a variety of electoral systems.

While individually well-designed, these projects and formats collectively suffer from a variety of problems:

- Adoption has been relatively limited
- There are few tools available to support the creation or consumption of the data
- The initiatives tend to be regionally focused and/or require significant technical skill in order to customize for international and regional differences in elections
- The data formats are largely intended to be produced and consumed by applications, rather than end-users, limiting use of the data to those with the technical skills to manipulate it.

Modelling Election Data

Based on the review summarised in the previous section, there are benefits in defining some new data formats to support the publication of open election data.

[17] https://votinginfoproject.org/
[18] http://popoloproject.com/
[19] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=election

There are some general requirements which can inform the scope and design of these formats. Specifically, formats should be:

- Easy for election officials to generate both manually, e.g. using simple spreadsheet tools, and automatically, e.g. through database exports or APIs
- Easy for both citizens and developers to consume using a range of tools, including simple spreadsheet applications
- Customizable and/or extensible to allow for international and regional differences in types of election data
- Flexible enough to support reporting of data at different levels of granularity, e.g. at country, region and administrative district levels
- Make reference to standard terms, code lists and reference data to ensure that data is clearly documented
- Focused on supporting transparency of electoral processes, rather than driving process automation.

The following sections present the design of a data model for election results and discuss the areas in which it must support customization for international uses.

More detail on the technical aspects of the model is provided in the accompanying technical specification and schemas.

Areas of Extensibility

Due to the variety of ways in which elections are conducted, a data format for elections will need to be extensible in a number of areas. The following sections highlight the main areas that need to be addressed.

Electoral System

One data point rarely referenced in existing election data is the electoral system under which the election has taken place. While election results are typically reported as a total number of votes or seats won, the means by which those totals were calculated can vary according to the type of electoral system used in the election. The families of electoral systems can be organised into a general classification[20].

These variations in election systems are important to recognise when performing comparative analysis of elections. This applies not just internationally, but also within a single country because

- the electoral system in use for a particular type of election may change over time
- different electoral systems may be used for different types of election

[20] http://aceproject.org/ace-en/topics/es/esd

- the use of hybrid election systems means that the same election may simultaneously use different systems in different regions

Categorising elections according to a broad family grouping, which is then referenced as a controlled vocabulary from election results, will help highlight differences and inform analysis.

Election Types

The types of elections for which data will be reported will vary according to the political system for the specific country or region.

For example in the UK there are parliamentary elections, European Parliament elections, local elections and mayoral and police commissioner elections.

Given the wide variety of different types of election, this is an area that will need to be customized on a per region basis.

Reference Data

As noted earlier in this report, there are some broad categories of reference data that are used to provide context to election results. These include:

- Political parties with which candidates may be associated
- Regions in which votes are reported, e.g. electoral wards or districts
- The administrative region in which a winning candidate may hold office

The structure of administrative and electoral regions will obviously vary across countries. However there are often national or international standards that can be used to provide common identifiers and reference data for this information.

Reference data on political parties should exist on a national level. For example the UK Electoral Commission provides a unique identifier for all registered parties. However this data may not always be available in a machine-readable form.

Wherever possible election data should make use of national or international standard identifiers for all key reference data.

Controlled Vocabulary for Voting Statistics

It is clear from reviewing existing approaches to election data publishing that there is some variety in the ways in which election results are reported.

Turnout is a measure of the number of eligible voters who cast a ballot at an election. Even this simple metrics can be difficult to calculate[21] and the numbers reported across countries may not be directly comparable. For example “eligible voters” may be an assessment of the size of the electorate or be restricted to just registered voters.

There are also different approaches to counting the votes used to calculate turnout, e.g:

- Total Participation Turnout -- counts all votes cast, whether valid or invalid and regardless of the method of voting
- Valid Vote Turnout -- counts only valid votes
- “Ballot box turnout” -- is defined by the UK electoral commission as being all votes cast at a polling station (whether valid or invalid) plus the number of valid postal votes

This variation is important to capture using a controlled vocabulary otherwise numbers may be incorrectly reported or compared.

Another key area of variation also relates to ballot counting. Votes are often divided into valid and invalid ballots. Ballots may be declared to be invalid for a number of reasons, e.g. they are blank or have been intentionally or accidentally spoilt[22]. The ACE project uses a slightly different definition of spoilt and rejected ballots[23] which separates out those ballots that have been spoiled but not added to the ballot box, from those which were spoiled and then rejected at the count.

Some election data provides a detailed breakdown of different types of invalid ballots, whereas other datasets just provide an overall figure. Again, this is an area where some controlled vocabulary is important.

Finally some election result reporting provides more than just headline figures for the number of votes cast: in some cases the method of voting is also indicated. Votes may be cast in a number of ways, e.g:

- in person at a polling station
- by proxy
- by post
- electronically, using an online application

The method of voting is often of interest for a number of reasons including: determining use and adoption of different methods by various demographics; if there are concerns about security of a particular voting method (e.g. an online submission); or, as noted above, in

[21] http://en.wikipedia.org/wiki/Voter_turnout#Measuring_turnout
[22] http://en.wikipedia.org/wiki/Spoilt_vote
[23] http://aceproject.org/ace-en/topics/vc/vce/vce02/vce02b

measuring turnout. Again, a controlled vocabulary would be useful to ensure clear reporting of data.

A Conceptual Model for Election Results

This section presents a simple conceptual model for describing election results. Accompanying this report is a more detailed technical specification that defines schema for both a tabular (CSV) and RDF view of election data. This section defines some of the key concepts that make up the conceptual model that guides the definition of these formats.

Election data can be divided up into two broad categories:

- **Reference Data** -- metadata about the election, its participants, polling stations, regions, etc
- **Statistical Data** -- election results and vote counts

Reference Data

Entity	Key Properties
Contest An election takes place on a specific date, or period of days. Contests may follow different rules that guide the number of votes and style of voting. Contests may take place for different reasons, e.g. a parliamentary election	• Name • Region • Start/end date of the election • Electoral type, e.g. parliamentary election • Electoral system, e.g. FPTP • List of Choices
Choice The choices presented to voters in a contest are typically a list of candidates and/or parties.	• Name of Party and/or Candidate presented as a choice
Candidate An option in a contest, usually a person running in an election	• Identifier • Name • Optionally, a party affiliation
Party A political organisation with whom a candidate may be associated	• Identifier • Name
Region A region in which a contest took place, or an administrative region in which a successful candidate holds office	• Identifier • Name • Optionally, for electoral districts the number of eligible voters

	• Optionally, parent/child relationships to describe an administrative region

Statistical Data

Election results data is inherently statistical, consisting of a measurement (e.g. number of votes) taken at a certain time and at a certain place. The W3C Data Cube vocabulary provides a conceptual model and an RDF vocabulary for describing statistical data:

- **Measures** -- the number being reported, e.g. the number of votes
- **Attributes** -- annotations that provide context to the measurement, e.g. units, or an indicator to note that the counted votes resulted in a win for a candidate.
- **Dimensions** -- contextual data that describe how the measure was collected, e.g. in which region, for which candidate, contest, etc. Dimensions values normally refer to a controlled vocabulary or fixed set of reference data, e.g. as described in the previous section
- **Observations** -- an individual measurement with a fixed set of dimensions and attributes. E.g. the number of valid votes reported in a specific region for a single candidate
- **Dataset** -- a collection of observations, e.g. the total results for an election or series of elections

The following table outlines the key measures, attributes and dimensions required to support election results

Component	Type	Description
Votes	Measure	Number of votes being reported
Electorate	Measure	Number of voters eligible to vote in an election
Seats	Measure	Number of seats allocated
Rank	Attribute	Ranking of choices candidates
Elected	Attribute	True or False. Indicates whether the counted votes resulted in a candidate being elected

Reporting Date	Attribute	The date in which the votes were reported. This may be different to the election date if votes are reported at a later date
Vote Category	Dimension	Controlled vocabulary describing categories of vote: e.g. valid, invalid, spoilt, rejected, etc
Voting Method	Dimension	Controlled vocabulary describing the method of voting, e.g. in-person, postal
Region	Dimension	The Region in which the votes are being reported. This is not necessarily the region in which a winning candidate may hold office
Contest	Dimension	The election contest in which the votes were counted

Collectively these components will support a variety of reporting styles and the ability to calculate a number of derived metrics, e.g. turnout, using a variety of criteria which can be defined by the individual data consumer or reporting body.

The benefits of using the Data Cube model as the framework for describing election results is that it can be easily extended to support new uses. For example to support analysis of voter foot-fall at specific polling stations two additional dimensions might be added allowing Observations to be reported in a more fine-grained, but still anonymised way:

- **Voting Period** -- period of the day, in which votes were submitted
- **Voting Station** -- the polling station at which the votes were submitted

The model could also be extended to support different types of contest:

- Opinion polls taken before elections might offer the same set of candidates but are collected and reported by a different organisation and have no official standing. Here the only difference is in who collects and reports on the results
- Referenda offer voters a choice between different options (e.g. joining or leaving the EU, or independence for Scotland). In this case instead of voting between candidates, voters are selecting between different choices. In this case the voter selection dimension would draw on a different controlled vocabulary.

Summary

Access to authoritative election data is an important part of a transparent democracy. Analysis of that data can help drive improvements to the democratic system by making it more efficient and increasing citizen engagement.

Currently while many governments publish election data it is rarely done so in a machine-readable format, making it difficult to re-use. An opportunity exists for improving this situation my making it easier for officials to publish open election data in simple formats.

This report has explored the various types of data available and has presented a simple conceptual model defining a common structure suitable for reporting election results.