development initiatives

Publish WhatYouFund
The Global Campaign for Aid Transparency

Identifying government entities

Exploration paper, September 2017, Joined-up Data Standards project

Tim Davies, Open Data Services Co-operative[1]

This paper explores options for **a universal method of identifying government entities** in datasets describing transactions from or to government agencies. It was developed based on dialogue with the org-id.guide partners and the IATI community between September 2016 and June 2017.

Contents

[1] This paper was written by Tim Davies of Open Data Services Co-operative with financial support from the Joined-up Data Standards (JUDS) project. JUDS is a joint effort of Development Initiatives (DI) and Publish What You Fund, with funding support from Omidyar Network. The JUDS team is grateful to Tim Davies for his work on this paper.

Introduction

Many of the datasets and standards created to further transparency created over recent years include information about transactions involving government entities. Persistent, non-overlapping and widely used identifiers are important to enable machines to combine data from multiple sources, and then to answer questions such as the following.

- How much money has been spent by the UK Department for Health with small or medium enterprises?

- How much money has the official Ugandan education system received from government donors in the last five years?

In these cases, the use of an identifier would allow human or machine data users to clearly identify: (a) the companies or charities in receipt of money; and (b) the government entities providing and receiving money. It is increasingly straightforward (though not entirely without challenge) to identify commercial organisations or charities unambiguously through the use of official identification and registration numbers. However, government entities often lack such stable public identifiers. As a result, there can be many missed connections in current datasets.

Previous efforts to identify a ‘Government Entity Identifier’ (GEI) for use in the International Aid Transparency Initiative (IATI) were not able to suggest a clear approach. This paper revisits the problem. Following a detailed exploration of the different requirements that any solution must address, we provide an updated assessment of potential methods to be adopted to better join up data about government agencies.

Mapping the challenge

The challenges of identifying government entities within datasets can be summarised with a simple example.

> A data publisher wants to record two transactions. The first transaction is a grant to a non-government organisation (NGO) in Bangladesh. The second transaction is a grant destined for the ‘Education Boards Bangladesh’, a unit of the ‘Ministry of Education’.
>
> For the transaction with an NGO, the publisher can look for bodies that identify or list NGOs in the country. From org-id.guide they will locate the NGO Affairs Bureau (code BD-NAB), and finding that the organisation they are funding has the identifier 0210 in the NGO Affairs Bureau list, can construct a unique identifier: BD-NAB-0210.
>
> For companies and NGOs there are generally one or more lists they ***must*** be on, in order to operate from, or in, a given country.
>
> However, for government agencies, there may be no such public lists. Furthermore, while drawing the boundaries between one NGO and another is straightforward, finding the dividing lines between parts of government is more complex. Should the legal entity in receipt of the transaction be listed as the ‘Government of Bangladesh’, the ‘Ministry of Education’, or ‘Education Boards Bangladesh’? The result is that the data publisher must either resort to an ad-hoc list like the OECD DAC’s Channel Codes, to use internal identifiers that will not map to the data published by another organisation, or publish only a plain-text organisation name.

The feasibility of different approaches to address this challenge depends upon the particular use-case for identifying government entities. Below, we explore three example use-cases to draw out particular requirements and challenges.

Use-case 1: Aid allocations and spending

Data publisher
A donor wants to publish information on aid allocations and spending with recipient government departments across 50 different countries.
Data users
A recipient government wants to align committed funding with departmental budgets. An aid analyst wants to understand the extent to which aid is being assigned to central government departments, or provided directly to sub-national regions and agencies.

The publisher would benefit from a common methodology that can be applied across all the countries to which it is providing funds.

The users in this case will benefit when: (a) different donors use the same identifier for the same government agency; (b) there is some additional meta-data attached to organisation identifiers, including the level of government.

Use-case 2: Contracting data

Data publisher
A government contracting portal needs to publish data about the departments issuing tenders and awards for goods and services. The portal collects information submitted by multiple agencies across the country, sometimes on their own behalf and sometimes aggregating information from multiple local procuring entities.
Data users
Government parties involved in a trade agreement want to generate statistics on the parts of government that have been making tenders available for international bids. A procurement watchdog organisation wants to monitor the tenders of a specific agency. An investigative journalist wants to explore the contracts issued by schools in a particular region of the country.

A central portal could create and publish data using its own internal identifiers. However, procurement data is often published at different levels of government, or the incoming data may, as noted, come from multiple agencies that are not clearly distinguished. This can lead to: (a) the same agency being assigned more than one identifier; or (b) different agencies being combined and described using the same identifier.

This case also highlights the challenges of delineating parts of the public sector. For example, a particular school may be fully within the state sector, legally structured as a part of an education authority, and given an identifier in a list of schools. Another school in the same area may be a free school or academy, linked to the state education system but with its own independent identity as a company or charity – and so with a number of different possible identifiers. A similar situation may occur with government-owned enterprises registered as companies but fully owned by the state.

Use-case 3: Freedom of Information requests

Data user
A Freedom of Information (FOI) submission system wants access to a regularly updated list of government agencies, and their contact information, for submission of FOI requests.

Governments change over time: departments are split or merged, and sub-units created or disbanded. Access to lists that keep track of these changes is desirable for many use-cases, including the FOI case described above.

Assessing the current situation

In this section we review some brief case studies exploring existing identification of government agencies within open datasets.

Identifiers in published data

International Aid Transparency Initiative

The International Aid Transparency Initiative (IATI) allows data publishers to specify the participating organisations for each aid activity. Using the @type attribute, they can indicate government parties to an activity.

Based on a November 2016 corpus of data, we found over 15,000 unique names against participating organisation elements identified as type ‘10’, or ‘Government’. The table below shows a random sample of the names given.

Organisation name[2]	**Reference/Identifier**
Belgian Development Cooperation	BE-10
BMZ	DE
Civilsamfund i Udvikling	(blank)
Dak Nong District Peoples Committee	(blank)
Danida	(blank)
DBP partner(s)	(blank)
Department for International Development	GB-1
DEPARTMENT OF PUBLIC WORKS AND HIGHWAYS	(blank)
DEPARTMENT OF TRANSPORTATION AND COMMUNICATIONS	(blank)
DFAT / Irish Aid	XM-DAC-21-1
DFAT / Irish Aid	XM-DAC-21-1
DFID	GB-1
Dutch Ministry of Foreign Affairs	NL-1
EU - European Development Fund	(blank)
Finland	FI
Foreign &amp; Commonwealth Office	GB-3
France	FR
Global Water Partnership	(blank)
Gouvernement du bÃƒÂ©nÃƒÂ©ficiaire	XM-DAC-12000
Gouvernement du donneur	XM-DAC-11000
Grundfos Holding A/S	(blank)
Health Centre of Matola Rio	(blank)
Institut for Menneskerettigheder	(blank)

[2] Data in the table is provided as contained within downloaded files.

International Development Association	(blank)
Japan	JP, XM-DAC-701
Japan International Cooperation Agency	XM-DAC-701-8
KERALA WATER AUTHORITY	(blank)
KOM partners	(blank)
LIAONING PROVINCIAL PEOPLE'S GOVERNMENT	(blank)
Ministry of Environment	(blank)
Ministry of Finance	(blank)
Ministry of Foreign Affairs	FI-3
Ministry of Foreign Affairs and Trade	NZ-1
Ministry of Foreign Affairs, Denmark	DK-1
Ministry of Planning and Investment	(blank)
MINISTRY OF PUBLIC WORKS AND TRANSPORT	(blank)
MINISTRY OF ROAD TRANSPORT AND HIGHWAYS	(blank)
MINISTRY OF URBAN DEVELOPMENT AND SACRED AREA DEVELOPMENT	(blank)
Netherlands Development Organisation	(blank)
Nordisk MinisterrÃƒÂ¥d	(blank)
Northwind Power Development Company	(blank)
Norwegian Church Aid	(blank)
OFFICE NATIONAL DES TELECOMMUNICATIONS	(blank)
Oxfam Ibis	(blank)
Patent- og VaremÃƒÂ¦rkestyrelsen	(blank)
Recipient government	(blank)
Service de CoopÃƒÂ©ration et d'Action Culturelle	FR-99
Servicio Nacional de Defensa Publica	(blank)
States of Jersey	(blank)
Swiss Agency for Development and Cooperation (SDC)	CH-4
The Ministry of Health of the Government of Mozambique	(blank)
THE PEOPLE'S GOVERNMENT OF HUNAN PROVINCE	(blank)
Uganda AIDS Commision	(blank)
United Kingdom	GB
United Nations Development Programme	(blank)

A review of this random sample, exploring the names and organisation references in use in IATI data, highlights a number of challenges.

1. **Use of organisation identifiers for government agencies is very limited**. And where identifier references are used, they are often not used correctly. For example, ‘GB’ is a country code, rather than an organisation reference.

2. **A number of non-governmental organisations are marked as ‘government’**. For example, ‘Oxfam Ibis’ appears with a type value of ‘10’, yet is clearly not a government agency.

3. **Several different levels of government are named:** from country-level declarations of donors (e.g. ‘United Kingdom’) to specific offices at subnational level (e.g. ‘KERALA WATER AUTHORITY’).

Contracts data

String matching

At the OpenGovToolbox hackathon in Paris in December 2016, OpenOpps.com took a sample of buyer names from central government procurement data in the UK and France. With a list of over 15,000 different strings representing buyer names, they applied a string-matching algorithm to the names. This resulted in approximately 60% of strings with a close match with at least one other string. This indicates:

1. That the same organisations had their names written in many different ways
2. The potential of string-matching processes to deal with at least some of this challenge – albeit without offering a guarantee, as such matching can lead to both false positives, and false negatives.

Identifiers available from government maintained lists

In initial research to update the list of organisation identifier lists held by the org-id.guide project, work has taken place to identify existing government-maintained lists of identifiers for government agencies, and to explore their characteristics. A small number of cases have been located so far, demonstrating a range of different approaches to managing government-agency identifiers.

United Kingdom: Government Single Domain (GB-GOVUK)

The Government Single Domain project in the United Kingdom has led to the creation of gov.uk with a page listing all central government organisations and their web pages at https://www.gov.uk/government/organisations. The list covers: ministerial departments (25), non-ministerial departments (21), agencies and other public bodies (376), high-profile groups (78), public corporations (10) and devolved administrations (3).

The GB-GOVUK prefix in the IATI Registration Agencies code list (inherited in org-id.guide) suggests using the last component of the organisation's URL to construct an identifier. This is a pragmatic move, but lacks a number of desirable features, such as the ability to understand change over time. If a department changes name, or is merged with another, the list at https://www.gov.uk/government/organisations will change. While the web team might use HTTP redirects to ensure links don't go dead, the goal will be to redirect people to relevant information, not to maintain a record of the way in which legal or financial responsibilities pass from one department to another.

Spain: Common Directory of Organizational Units and Offices (ES-DIR3)

The Common Directory of Organizational Units and Offices (DIR3) is a project to improve interoperability between public administration units in Spain. Within this, a list of all public bodies is maintained by the Centro de Transferencia de Tecnología (Technology Transfer Centre) as a number of regularly updated Excel and RDF files are provided for download. There does not appear to be any information provided on change over time of government entities that would support reconciliation of data after an entity is renamed.

Netherlands: Overheid.nl (NL-OWMS)

Overheid.nl is the central access point to all information about government organisations of the Netherlands. The Overheid.nl Web Metadata Standard (OWMS) is the metadata standard for information from the Dutch government on the Internet. It contains URIs for a wide range of government bodies, including national, local and regional government and water boards.

It provides a linked open dataset which contains ontological information about the relationship between those organisations (e.g. listing parent agencies, or noting organisations that succeed previous organisations). A management plan is in place for updating the information.

United Kingdom: Edubase / Schools Register (GB-EDU)

All educational establishments in the United Kingdom must be registered with the Department of Education. Edubase is provided by the Department of Education and provides a Unique Reference Number for each school, university and other educational establishment in England and Wales.

Edubase is currently in the process of being replaced by a new register, produced by the Open Registers project. This is a government initiative to create stable registers of core concepts and entities across government.

Czech Republic: Access to Registers of Economic Subjects (CZ-ICO)

The Access to Registers of Economic Subjects system (ARES) is an information system which collates data from several public registers in the Czech Republic, including public registers comprising: the Commercial Register, Federal Register, the Register of Foundations, Register's Institute, Register of Public Service Companies, Trade Register, and the Register of Economic Entities. Some government entities can be found from searching the register, although the scope and comprehensiveness of the register is unclear.

In our scoping conversations, we were directed to government charts of accounts, or counter-party identifier datasets, as a possible source of identification lists of government organisations. However, we were able to locate only the UK Whole of Government Accounts Counter Party ID list as an example of these. In this list, multiple organisations are grouped as a single counter-party. While the spreadsheets for each year maintain a basic change-log, they do not have comprehensive meta-data. The lists also appear to include non-governmental entities (e.g. trusts in receipt of public funds), without indicating the legal status of these organisations.

From the examples above, our use-cases, and scoping research, three key issues arise for consideration.

1. **It is important to understand the scope of each list**. Some lists may aim to cover all of government, others only central government, others a particular kind of entity (e.g. schools). A national list is likely to stop at the level of a local authority; but some use-cases might want to identify departments and entities within a particular local authority.

2. **Lists range from comprehensive registers to ad-hoc authority lists**. A register aims to cover all the entities of a particular kind. An ad-hoc authority list will add new entries as required. For example, if an authority has never been involved in contracting, it might not be in a register of economically active public authorities.

3. **Governments change over time.** Departments are created, renamed, reshaped, merged and closed. It is desirable for lists to maintain old identifiers, and keep information on what happened to a particular department or agency (e.g. where its responsibilities were transferred to a new entity).

Existing standards for government entity information

The **Core Public Organisation Vocabulary** (CPOV) of the European Commission builds on the World Wide Web Consortium (W3C) Organization Ontology, and provides a common set of terms for describing organisations. It includes fields for identifiers, although this does not reference any specific identifier scheme, nothing that: "Many organizations are referred to by an acronym or some other identifier."

CPOV provides terms for describing the spatial coverage of a public organisation, and its purpose, using classifications such as Classification of the Functions of Government (COFOG) or some other controlled vocabulary. CPOV also includes properties for relating organisations, with hasUnit/unitOf and hasMember/memberOf relationships. See https://joinup.ec.europa.eu/asset/cpov/home for further details.

Requirements

Based on the analysis above, we can distil 11 core requirements for a future identifier system, against which any solution should be tested.

An identifier system should:

1. Allow identification of central government departments
2. Allow identification of local government units
3. Allow identification of government agencies
4. Allow identification of bodies such as schools or hospitals
5. Provide access to information on the level of government the body operates at
6. Provide access to information on the history of the body (e.g. bodies it replaced)
7. Provide access to information on the hierarchical position of the body
8. Uniquely identify a single entity (e.g. two distinct government entities should not share a code)
9. Provide persistent identifiers that change only when the integrity or legal status of the identified organisation changes
10. Be appropriate for use in a single country
11. Be appropriate for use across countries (for example, identifying all the mining ministries around the world).

These requirements may be prioritised differently by different use-cases.

Candidate approaches

In this section we consider different possible approaches to these challenges.

Distributed identifier lists

Maintaining the current org-id.guide approach would involve seeking to locate existing lists of government entity identifiers. This could be pursued through either a campaign for governments to maintain their own registers or a more ad hoc approach of continued research to locate sources of identification at the point when they are needed.

A campaign for government-maintained registers

Through fora such as the Open Government Partnership, a concerted campaign could call for governments to maintain clear registers of their organisational entities.

An ad-hoc approach

Experience suggests that, with enough digging, a list of government entities in a given country can be found. These may not be very high-quality lists, but they may be good enough to meet the needs of basic use-cases.

Enhanced identification information

Instead of trying to find identifier strings which can be one-for-one matched to determine that two sources of data are referring to the same organisation, this approach focuses on capturing additional meta-data that can be used in both direct analysis and processes of de-duplication. This essentially places the burden of describing the organisation onto the data publishers, rather than assuming that descriptive information about the organisation can be derived from the identifier itself.

As meta-data

Capturing this additional information as meta-data would involve agreeing a common set of fields and code-lists for extra information. At its simplest, this may include asking that publishers always pair either an address (including country) or a URL with the organisation name. These two additional pieces of information could be fed into matching algorithms to help avoid false positives, and reduce the number of false negatives when carrying out automated matching based on organisation names.

Other candidate meta-data fields include:

- A **purpose** code to describe the focus of the government entity, drawing upon an established vocabulary such as COFOG
- Relevant **spatial** information for the extent of the jurisdiction of the government entity
- A code to indicate the **level** of government of the entity.

Example:
If a standard provided the following fields (according to a defined standard), the chances of incorrect matches are substantially reduced.

Name	Jurisdiction	Purpose (COFOG)	Spatial coverage	Level	Address	URL
Department of Education	GB	09	ADM1	National	London, UK	http://...

This approach was considered but rejected by the IATI Technical Advisory Group (TAG) Standards Day 2017 as it was felt that the multiple fields added a level of complexity for both publishers and data users.

Embedded in the identifier

Finding a natural way to embed this information in an identifier string is difficult. One option would be to introduce a parenthetical component to the end of an identifier string to contain: (COFOG code / administrative level / address information).

Any tool using the data could then try a strategy of ID reconciliation based on:

- First matching the non-parenthetical component of the identifier
- Then, if this does not yield a match, trying to match the organisation name, subject to the constraints in parentheses.

For example, the education department at Oxfordshire County Council might be identified as:

- OCC Education Dept, XI-PB-gb/oxfordshire-county-council (09/ADM2/Oxfordshire), *and*
- Oxfordshire Education Services, XR-NUTS-UKJ14 (09/ADM2/OX1 1ND).

In these cases, a direct match on identifiers or names would fail, but the information in the parentheses indicates that both of these are education organisations, working at the County level, in the Oxfordshire geographical area. While this does not give a conclusive match, it can aid matching.

A single identifier source

There are a number of candidate platforms or approaches that could be adopted to develop a single hub of identifiers. For example:

- The Public Bodies (http://publicbodies.org/) project set out to create a URL for every part of government, based on a simple GitHub repository which can accept user contributions with a list of government entities, and an assigned URL for each one.

- The EveryPolitician project (http://everypolitician.org/) builds on this model, but for politicians rather than government entities, using a mix of user-contributions and scrapers to keep information up to date. This approach (or scraping data into a central register, as well as accepting manual submissions) could be applied to a platform in the style of Public Bodies.

- The Thomson Reuters PermID platform (https://permid.org/) assigns identifiers to entities based on Thomson Reuters own data knowledge base. There is no clear pathway for user contributions to this dataset when data is missing or mis-classified.

- WikiData includes identifiers for a wide range of concepts, including government entities. For example:

  this query:

```
SELECT ?item ?itemLabel ?class ?classLabel
WHERE
{
   ?item wdt:P31/wdt:P279* wd:Q2659904.
   ?item wdt:P17 wd:Q145.
   ?item wdt:P31 ?class.
    SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }

} LIMIT 10
```

  searches for government entities in England known to Wikidata. Wikidata will host information on any entities that pass a notability threshold (which most national and regional government entities should easily pass), and provides clear pathways for user contribution.

Each of these platforms provides data under an open licence. However, none currently offers comprehensive coverage of the government entities according to the requirements specified above.

Proposal: a hybrid approach

Based on discussion on the above options with the IATI TAG, org-id.guide partners and the wider community, this paper proposes a hybrid approach. This is based on:

- **Maintaining the existing distributed organisation identifier list approach** paired with a focused call for governments to share details of existing organisation identifier lists they maintain.
- **Creating a ‘list of last resort’** through which publishers can register temporary identifiers for government entities with no existing identifiers.

This list of last resort will store detailed meta-data for each user-contributed entry covering organisation type, function (using COFOG for government entities), area of jurisdiction (using Geonames identifiers for ADM levels), website and head-office location. It will include space for alternative names (AKAs) to be recorded, and for alternative identifiers to be included once known. The list will be registered in org-id.guide as ZZ-TMP.

A software tool will be required to manage this list, accepting user contributions, and providing a lightweight moderation mechanism. This tool should:

- Allow a search of cached organisation information by country, type, sector and name
- Allow creation of new identifiers in the ZZ-TMP namespace
- Be embedded within org-id.guide and accessed in cases where no primary or secondary lists are available.